



Contributions à l'extraction de connaissances à partir de données biologiques

Malika Smaïl-Tabbone

► To cite this version:

Malika Smaïl-Tabbone. Contributions à l'extraction de connaissances à partir de données biologiques. Apprentissage [cs.LG]. Université de Lorraine, 2014. tel-01093943v2

HAL Id: tel-01093943

<https://inria.hal.science/tel-01093943v2>

Submitted on 23 Jan 2015

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

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

Contributions à l'extraction de connaissances à partir de données biologiques

THÈSE

soutenue le 14 novembre 2014

pour l'obtention d'une

Habilitation de l'Université de Lorraine
(mention informatique)

par

Malika SMAIL-TABBONE

Composition du jury

Rapporteurs : Christine FROIDEVAUX, Professeur Université Paris-Sud
Céline ROUVEIROL, Professeur Université Paris-Nord
Julie THOMPSON, Directeur de Recherche CNRS

Examineurs : Anne BOYER, Professeur Université de Lorraine
Bruno CREMILLEUX, Professeur Université de Caen
Marie-Dominique DEVIGNES, Chargée de Recherche CNRS

Mis en page avec la classe thesul.

Table des matières

Avant-propos	9
1 Introduction et contexte	11
1.1 Introduction	11
1.2 Analyse des données biologiques : spécificités et enjeux	12
1.2.1 Les bases de données biologiques : du volume et de la diversité	12
1.2.2 Attentes des biologistes	14
1.3 Extraction de Connaissances à partir de Données : tout un processus	16
1.4 La fouille de données	17
1.4.1 Extraction de motifs et de règles d'association	18
1.4.2 Classification supervisée	20
1.4.3 Classification non supervisée (clustering)	21
1.4.4 Analyse formelle de concepts	22
1.5 Plan du document	24
2 Préparation pour la fouille des données biologiques	25
2.1 Introduction	25
2.2 Organisation du ressourceome biologique et découverte de sources de données	25
2.3 Intégration des données biologiques	27
2.4 Contribution	28
2.4.1 BioRegistry : Organisation et découverte de bases de données biologiques	28
2.4.2 MODIM : Intégration de données fondée sur un modèle en vue de la fouille de données	31
2.4.3 Mesure de similarité sémantique entre objets biologiques	34
2.5 Conclusion	36
3 Fouille de données biologiques relationnelles	39
3.1 Introduction	39
3.2 La fouille de données relationnelles	39

3.2.1	Apprentissage de Concept par Programmation logique inductive	40
3.2.2	Le système ALEPH	42
3.2.3	Extension (<i>Upgrad</i>) de méthodes de fouille de données classiques	43
3.2.4	Éléments sur les performances des systèmes de fouille de données relationnelles	44
3.3	Fouille de données relationnelles appliquée aux données biologiques	44
3.4	Contribution	45
3.4.1	Caractérisation de sites tridimensionnels d'interactions protéine-protéine .	46
3.4.2	Définition et caractérisation de profils d'effets secondaires de médicaments	48
3.4.3	Environnement logiciel pour le déploiement de la fouille de données relationnelles	51
3.5	Conclusion	52
4	Aide à l'analyse et à l'interprétation des résultats de la fouille	53
4.1	Introduction	53
4.2	Évaluation, validation et visualisation des résultats de la fouille de données . . .	53
4.3	Approche LeGo : des motifs locaux aux modèles globaux	54
4.4	Bases de données inductives	55
4.5	Contribution	55
4.5.1	L'analyse formelle de concepts comme outil d'interprétation d'une théorie	57
4.5.2	Combinaison "à la LeGo" de plusieurs méthodes de fouille au service d'un problème d'apprentissage	58
4.5.3	Conclusion	59
5	Bilan et perspectives de recherche	63
5.1	Bilan	63
5.2	Projet de recherche	64
5.2.1	Prédicteurs en aval de la programmation logique inductive	64
5.2.2	Introduction de la transduction dans le processus d'ECD	65
5.2.3	Intégration des données biologiques ouvertes et liées dans le processus d'ECD	67
5.2.4	Des données patients aux connaissances : vers la médecine personnalisée .	69
5.3	Conclusion	71
	Références bibliographiques de l'auteur	73
	Références bibliographiques	81

Dossier de présentation	97
1 Diplômes universitaires	97
2 Expérience professionnelle et statut actuel	97
3 Résumé de mon activité de recherche	97
4 Activités d'encadrement	100
5 Responsabilités administratives : participation aux conseils et mandats nationaux	104
6 Résumé des activités pédagogiques	104
7 Principales publications au 1er avril 2014	105

Remerciements

Je voudrais tout d'abord adresser mes remerciements les plus chaleureux à Marie-Dominique auprès de qui j'ai beaucoup appris tant sur le plan de la biologie que sur le plan humain. Grâce à toi, j'ai découvert le monde de la recherche en biologie et ses codes. Je n'ai pas vu passer ces années post-thèse tant notre activité a été riche et motivante. C'est un privilège et un grand plaisir de travailler avec toi et de partager notre bureau.

J'exprime ma gratitude aux membres du jury qui ont pris le temps de juger ce travail.

Une pensée amicale et un grand merci à :

- Amedeo Napoli pour m'avoir donné l'opportunité de faire partie des orpailleurs
- à Karl Tombre qui m'a permis de bénéficier d'une délégation lorsqu'il était directeur d'Inria Grand-Est,
- à tous les orpailleurs notamment Adrien Coulet, Jean Lieber, Bernard Maigret, Chedy Raïssi, Dave Ritchie, Yannick Toussaint,
- aux collègues biologistes, micro-biologistes, cliniciens, automaticiens, statisticiens avec qui j'ai eu le plaisir de collaborer. Je citerai, en oubliant sûrement, Eliane Albuisson, Taha Boukhobza, Jean Devignes, Lionel Domenjoud, Hélène Dumond, Aurélie Gueudin-Muller, Nathalie Leblond, Valérie Leclère, Olivier Poch, Maude Pupin
- mes collègues de l'ex-équipe EXPRIM : Marion Créhange, Odile Thiéry, Odile Foucaut, Géral Duffing
- mes collègues de la sphère enseignement : Nacer Boudjlida, Isabelle Debled, Lotfi et Nadia Bellalem, Adrien Coulet, Bertrand Aigle, Nathalie et Pierre Leblond, Brigitte Wrobel-Dautcourt, Brigitte Jaray
- aux docteurs ou doctorants avec qui j'ai (eu) le plaisir de travailler : Gérald Duffing, Nizar Messai, Adrien Coulet, Saliha Yilmaz, Léo Ghemtio, Anisah Ghoorah, Sid-Ahmed Benab-derrahmane, Emmanuel Bresso, Mehwish Alam, Gabin Personeni
- aux étudiants stagiaires qui ont contribué à divers projets
- aux ingénieurs jeunes diplômés (Birama Ndiaye et Renaud Grisoni) et aux ingénieurs CNAM (ou pas), André Schaaff, Philippe Franiatte, et Laurent Pierron
- à tou(te)s mes étudiant(e)s passé(e)s et à venir

Last but not least, une dédicace spéciale à mon super mari(o) pour son soutien indéfectible et à nos quatre enfants (merveilleux, chacun à sa manière). Je forme le vœu qu'ils trouvent leur voie sans embûche...

A mon (super) Mari(o)
A ma famille

Avant-propos

Ce document résume le fruit de mes recherches menées principalement dans l'équipe Orpailleur du LORIA. Depuis l'année 2000, ces recherches ont porté sur l'intégration de données biologiques puis de façon conjointe sur la fouille de ces données en vue de l'extraction de connaissances. Le premier objectif de ce document est de synthétiser ces recherches dans un cadre cohérent et de les positionner par rapport à l'état de l'art contemporain afin de faciliter la lecture des articles publiés adjoints à ce document. Le second objectif est de dégager les principales directions d'un projet de recherche dans la continuité naturelle de ces travaux. La partie Dossier de présentation détaille mon *Curriculum vitae* à travers mon parcours professionnel et une synthèse des mes activités de recherche, d'encadrement, d'administration, et d'enseignement.

Les travaux présentés dans ce document sont également le fruit de mon activité de co-encadrement et de collaboration scientifique. Bien que je ne mentionne pas systématiquement le nom des différentes personnes qui ont contribué à chaque partie, cela apparait dans la signature de nos publications communes et l'usage fréquent de la première personne du pluriel est là pour souligner que ces résultats sont partagés avec ces personnes avec qui j'ai eu le plaisir de travailler incluant mes collègues chercheurs, des doctorants, des stagiaires et des ingénieurs.

Afin de faciliter la lecture du document, les références aux publications dont je suis co-auteur sont citées selon un mode numérique (e.g [8]) tandis que les autres le sont selon un mode alphanumérique (e.g [Yan 2005]).

Chapitre 1

Introduction et contexte

1.1 Introduction

L'histoire relativement récente de la bioinformatique est très liée à celle de la biologie moléculaire puisqu'elle couvre principalement les moyens informatiques de stockage et d'analyse des données relatives aux molécules du vivant ou bio-molécules. Dans ce contexte, la maîtrise et l'exploitation des ressources numériques est aujourd'hui un défi constituant l'un des enjeux majeurs de la biologie du XXIème siècle [31]. En effet, les techniques à haut débit en biologie expérimentale ont conduit à une multiplication des ressources numériques sous forme de bases de données qui servent à structurer et à entreposer les résultats des expériences mais aussi de programmes qui permettent de traiter ces données et d'en produire des nouvelles ou qui analysent la littérature à la recherche de données factuelles. . . Ces gisements de données sont disponibles au plus grand nombre et ne demandent qu'à être exploités. Ainsi, les biologistes ont d'énormes espoirs quant à l'analyse de ces données afin d'en extraire des connaissances qui leur permettront de résoudre divers problèmes tels que l'identification de gènes responsables de maladies, la caractérisation de médicaments du point de vue de leurs effets secondaires indésirables, la prédiction d'interactions entre bio-molécules. . .

En parallèle, de nombreux travaux menés dans la communauté de fouille de données et d'apprentissage automatique ont produit une large panoplie d'algorithmes efficaces pour réaliser différentes tâches d'apprentissage appliquées à des données réelles. Je ne ferai pas, dans ce document, une distinction rigoureuse entre le concept d'*apprentissage automatique* et celui de *fouille de données*. La définition de la fouille de données que je retiens est la mise en pratique des outils et des techniques de l'apprentissage automatique.

Néanmoins, la fouille de données se révèle difficile étant donné la complexité des données, leur hétérogénéité mais aussi leurs imperfections. Il apparaît clairement que des solutions méthodologiques et pragmatiques sont à trouver pour faciliter l'analyse de ces données. Ces solutions devront s'appuyer sur le processus itératif plus complet d'*Extraction de Connaissances à partir de Données* (ECD) englobant la fouille de données mais aussi les étapes cruciales (en amont) d'intégration et de préparation des données et les étapes décisives (en aval) d'évaluation et d'aide à l'interprétation des résultats. Une question importante porte sur les possibilités d'exploitation des connaissances du domaine dans le processus d'ECD menant à l'ECD guidée par la connaissance du domaine[LNST06]. Le processus devrait ainsi bénéficier des connaissances déjà établies afin d'extraire des connaissances moins triviales.

Mon activité de recherche et de co-encadrement est centrée sur cette problématique et s'est déroulée dans le cadre de plusieurs collaborations et projets d'extraction de connaissances à partir de données biologiques pour la plupart. Les contributions de cette activité peuvent être déclinées selon les trois étapes de l'extraction de connaissances à partir de données relationnelles :

1. Intégration de données dirigée par un modèle relationnel de données : modéliser un sous-domaine d'intérêt, collecter et intégrer des données distribuées en étant guidé par le modèle établi, et si nécessaire réduire la dimensionnalité des données (Chapitre 2).
2. Fouille de données complexes en facilitant la mise en œuvre de diverses méthodes permettant de découvrir à partir de données relationnelles notamment des règles logiques (exprimées en logique du premier ordre) grâce à la programmation logique inductive. En effet la formidable complexité des données biologiques s'accommode mal d'un aplatissement dans une table unique et il s'agit ici de permettre la fouille des données telles qu'elles ont été modélisées et intégrées (Chapitre 3).
3. Aide à l'évaluation et l'interprétation des modèles trouvés en permettant leur persistance et l'itération du processus d'ECD éventuellement précédée d'un changement de perspective sur les données (Chapitre 4).

Ce cadre conceptuel s'est imposé au fur et à mesure de notre activité de recherche et de développement et nous a conduit à proposer des éléments de méthodologie pour l'extraction de connaissances à partir de données complexes guidée par les connaissances du domaine. Nous avons en outre défini une mesure de similarité sémantique entre objets biologiques (gènes, protéines, activités de médicaments) fondée sur une ontologie de domaine qui permet de grouper des objets d'une part et de réduire la dimensionnalité d'ensembles de données d'autre part. Ainsi, notre cadre d'ECD permet d'exploiter des ontologies biologiques disponibles pour la préparation des données mais aussi pour la fouille de données et pour l'aide à l'interprétation des résultats.

Afin de présenter le contexte des travaux décrits dans ce document, je synthétise dans un premier temps les particularités des données biologiques et de leur analyse (Section 1.2). Je présente ensuite le modèle de processus d'ECD (Section 1.3) et de façon un peu plus détaillée l'étape de fouille de données en décrivant les principes des méthodes dédiées aux principales tâches de fouille (Section 1.4). Le plan du reste du document clôturera ce chapitre.

1.2 Analyse des données biologiques : spécificités et enjeux

L'accroissement continu des ressources numériques depuis les années 90 est certainement une chance pour la recherche en biologie. Cependant, il pose de nombreux défis que je tenterai de résumer après la description des spécificités des données biologiques.

1.2.1 Les bases de données biologiques : du volume et de la diversité

Les premières données biologiques mises à disposition sont les séquences nucléiques et les séquences protéiques à travers les bases GenBank, EMBL, DDBJ, et Swissprot. Il est commun aujourd'hui de décrire la croissance des banques de séquences par l'intermédiaire d'une courbe exponentielle. En plus du volume croissant des données disponibles, nous notons surtout une très grande diversification dans la nature de ces données.

Type de données	BDs	Nature des BDs
Protéines connues	Uniprot	-
Structures tri-dimensionnelles de protéines	PDB	-
Interactions protéine-protéine	IntAct, SCOPPI, 3DID	Secondaires
Gènes connus	NCBI Gene	-
Réseaux biologiques connus	KEGG PATHWAYS, BioCARTA	-
Informations sur les médicaments	DrugBank	-
Effets secondaires connus de médicaments	SIDER, STICH	-
Variants génomiques connus pour les gènes	dbSNP	-
Transcription de gènes	GEO (NCBI), ArrayExpress (EBI)	-
Maladies génétiques humaines et gènes responsables	OMIM	Spécialisée pour l'Homme
Diverses données sur un Organisme Modèle (OM)	FlyBase, MGI, WormBase	Intégrées et Spécialisées pour un OM
Littérature des principales revues bio-médicales	MEDLINE	Secondaire

TABLE 1.1 – Quelques types de données et des bases de données associées

Les principaux facteurs de cette croissance et de cette diversification sont l'arrivée du web qui a banalisé l'accès à l'information via le réseau Internet d'une part, et d'autre part l'essor des technologies dites à haut débit permettant l'analyse parallèle et massive d'échantillons biologiques. Bien que coûteux, ces équipements et méthodes s'imposent petit à petit comme des références et déchargent ainsi le chercheur de tâches répétitives. L'un des résultats visibles est la mutualisation des installations sous forme de plateformes. Moins visible mais tout aussi importante, l'accumulation progressive des données suscite une sollicitation croissante de traitements informatiques capables de les structurer et de leur donner un sens. Le tableau 1.1 énumère quelques types de données biologiques et des bases de données correspondantes. Depuis 2000, une compilation des meilleures bases de données de biologie moléculaire (*Molecular Biology Database Collection*) publiée par M. Y. Galperin dans le numéro de janvier de la revue *Nucleic Acids Research* fournit une information fiable sur l'existant [FSG13]. Ainsi, le nombre de bases de données (de couverture et de qualité significatives) est passé de près de 300 en 2003 à plus de 1500 en 2013.

Essai de classification par rapport au contenu Bien qu'il soit difficile d'établir une classification consensuelle des bases de données biologiques, quelques distinctions peuvent être faites. On peut par exemple distinguer des bases de données *généralistes* telles que les banques de séquences toutes espèces confondues (GenBank, Uniprot), des bases de données *spécialisées* par rapport à une espèce (MGI pour la souris, FlyBase pour la drosophile) ou à un type de données (Eukaryotic Promoter Database pour les promoteurs de transcription des gènes chez les eucaryotes). De nombreuses bases de données peuvent être qualifiées de bases de données *secondaires* parce qu'elles enrichissent des données contenues dans des sources *primaires* telles que les bases de données de séquences. C'est le cas par exemple de la base de données PROSITE qui utilise les séquences protéiques des familles de protéines pour en extraire des motifs (suites d'acides aminés) conservés. Par ailleurs, certaines bases de données spécialisées ou secondaires sont dites *intégrées* car leurs données proviennent de plusieurs sources. La base de données InterPro, maintenue à l'EBI, est un exemple de base de données intégrée très riche et généraliste puisqu'elle vise à recenser tous les domaines protéiques caractérisés à ce jour en faisant les correspondances entre les motifs et domaines recensés par plus d'une quinzaine de sources telles que ProSite, Pfam, ou PRINTS. Chaque entrée de la base InterPro fait aussi référence aux entrées UniProt des protéines contenant ce domaine. En règle générale, la construction d'une base de données intégrée est l'occasion de réaliser un nettoyage des données conduisant à une information de meilleure qualité et de restructurer les données selon un modèle cohérent tenant compte des connaissances du domaine et facilitant l'exploitation des données.

Divers formats et modèles de données Au départ, les banques de données biologiques étaient de simples fichiers plats mais assez rapidement la généralisation des systèmes de bases

de données relationnelles a fait que les biologistes, du moins les organismes ou institutions mettant à disposition des données et chargés de leur maintenance, se sont orientés vers des modèles relationnels afin de minimiser redondance, risques d'erreurs, et anomalies de mise à jour. Quel que soit leur format, les bases de données biologiques ont un contenu semi-structuré. En effet les données contiennent à la fois des champs plus ou moins atomiques (dates, mots clés, identifiants) et des champs textuels. MEDLINE en est un exemple avec de nombreux champs de méta-données et d'indexation du contenu d'un article à l'aide du thésaurus MeSH et le champ *abstract* correspondant au résumé de l'article.

Depuis la fin des années 90, le langage XML (eXtensible Markup Language) s'est imposé pour les données biologiques semi-structurées. XML permet de définir pour un domaine ou un type de document particulier un ensemble de balises et une syntaxe adaptés. Grâce à son expressivité et aux nombreux outils et langages associés, XML s'est imposé comme un format d'échange entre bases de données [AVB01]. Les outils disponibles (XPath, XQuery) permettent alors à l'utilisateur de sélectionner une partie des données du document de sortie et de les combiner aux données extraites d'une autre base de données, par exemple. Les dernières années ont vu l'arrivée de langages du web sémantique tels que RDF pour lesquels une syntaxe XML a été définie. Nous reviendrons sur le rôle du langage RDF dans la description des méta-données et dans l'essor des données ouvertes et liées (Chapitres 2 et 5)

Divers modes d'accès Les bases de données biologiques sont mises à disposition et maintenues par de nombreux organismes (e.g., EMBL-EBI, NCBI). Par conséquent, les interfaces homme-machine (IHM) d'interrogation sont spécifiques à chaque organisme voire à chaque base de données exigeant du biologiste un effort d'adaptation. Certaines initiatives ont eu pour objectif de faciliter l'accès à diverses bases de données à l'aide d'une IHM unique. C'est le cas des projets SRS et BioMart¹ développés à l'EBI (European Bioinformatics Institute) [ZLAE02, HBS⁺09]. A cela s'ajoutent des interfaces de programmation fondées sur le protocole SOAP destinées à faciliter l'interopérabilité des ressources biologiques qu'elles soient sources de données ou programmes d'analyse. Ainsi, le projet Taverna/myGrid² a pour objectif d'établir des librairies de services web donnant accès à diverses ressources biologiques et de permettre au chercheur de définir et de partager des scénarios (ou workflows) d'analyse enchaînant des services [MSRO⁺10].

1.2.2 Attentes des biologistes

L'accroissement et la diversification des données disponibles ont profondément modifié la démarche des chercheurs en biologie. Le temps de la trilogie *une hypothèse, une expérience, un article* est bien révolu. Certains scientifiques parlent même de recherche pilotée par l'ignorance tant les nouvelles possibilités se prêtent à des explorations tous azimuts. Même à l'échelle d'un petit laboratoire ou d'une petite équipe, le recours aux ressources numériques est devenu incontournable. Au moment où le biologiste pose un nouveau problème ou se pose une question, en plus de la revue de la littérature, il doit désormais s'assurer que la réponse n'est pas déjà disponible dans une des bases de données publiques. Il est alors confronté à un premier problème : identifier quelle(s) base(s) de données sont susceptibles d'être intéressantes pour sa question. Cela correspond à un problème de découverte de ressources qui doit être automatisé vu le nombre croissant de ressources disponibles.

1. <http://www.biomart.org/biomart/martview>

2. [urlhttp://www.mygrid.org.uk](http://www.mygrid.org.uk)

Une fois les ressources pertinentes identifiées, si la question posée n'est pas ponctuelle (exemple : quels domaines composent une protéine ? quels gènes sont responsables d'une maladie ?) se pose le problème de collecter l'ensemble des données d'intérêt à partir de différentes sources plus ou moins en accord sur les mêmes sujets. Cela correspond au problème d'intégration de données. Dans le cas où la réponse à la question posée n'est pas déjà présente dans les bases de données, une exploration expérimentale peut s'avérer nécessaire et les technologies à haut-débit permettent de réaliser des expériences à large échelle. Par exemple, le biologiste qui explore l'effet d'un médicament précis sur les patients qui souffrent d'un type de cancer peut en une seule expérience analyser la transcription³ de milliers de gènes de patients dans différentes situations (dans différents tissus, en association à un autre médicament, à différents stades de la maladie...). A ce stade, le biologiste est confronté à la difficulté d'interpréter les gros volumes de données plus ou moins brutes qui résultent de son expérience. Des analyses automatiques de ces données sont alors nécessaires. Certaines analyses statistiques sont fournies sous forme de routines sur certaines plateformes mais le biologiste souhaite souvent aller plus loin et confronter tout ou partie de ses résultats expérimentaux avec d'autres résultats ou alors remettre ses données en contexte en les complétant avec par exemple les annotations connues sur les gènes (leurs processus biologiques, leur localisation dans la cellule...). Se repose donc le problème d'intégration de données en amont d'une itération d'étapes de fouille de données. Pour chaque expérience de fouille de données, le biologiste a alors besoin d'assistance pour définir le problème de fouille, sélectionner et préparer ses données mais aussi interpréter les résultats de la fouille. Nous reconnaissons ici le processus d'Extraction de Connaissances à partir de Données (ECD).

En résumé, les biologistes ont d'énormes attentes lors de la recherche et de l'exploitation de leurs données, que ces données soient produites expérimentalement ou collectées à partir de sources publiques. Dans ce contexte, trois facteurs viennent renforcer la pression sur les informaticiens ou les bio-informaticiens :

- la complexité des données biologiques intégrées nous oblige à explorer diverses méthodes de fouille de données. En effet, les problèmes d'analyse posés sont divers et les modèles de données comportent souvent une forte composante relationnelle et structurelle qu'il est important de prendre en compte. Le volume des données pousse également à sélectionner les algorithmes de fouille de données capables de produire des résultats à l'échelle des données réelles ;
- la connaissance biologique se structure petit à petit sous forme d'ontologies lesquelles, comme les bases de données dans les années 90, sont mises à disposition et peuvent être utilisées à différents stades du processus d'analyse de données biologiques. L'exemple d'ontologie le plus connu est sans doute Gene Ontology qui permet de structurer le vocabulaire d'annotation des produits des gènes selon trois aspects : les processus Biologiques (BP) auxquels ils participent, les fonctions moléculaires (MF) qu'ils remplissent, et les composants cellulaires (CC) dans lesquels ils peuvent se trouver [Con10]. Deux autres exemples sont les vocabulaires structurés MeSH utilisé pour l'indexation des articles scientifiques (méta-données mais aussi contenu) et MedDRA pour décrire l'action des médicaments et des vaccins ;
- fouiller des données biologiques et présenter au biologiste un nombre important voire pléthorique de motifs (que l'on peut définir de façon informelle comme des régularités observées dans les données) n'est pas satisfaisant et ne peut pas être la fin du processus d'analyse.

3. La transcription correspond au fait qu'un gène produit de l'ARN en quantité variable, ARN qui donnera ensuite lieu à une protéine.

1.3 Extraction de Connaissances à partir de Données : tout un processus

L'extraction de connaissances à partir de données (ECD) est un processus multi-étapes, itératif et interactif [FPSS96]. L'étape centrale est évidemment celle de la fouille de données mais elle est précédée et suivie d'étapes importantes :

1. Intégration des données : cette étape sert à collecter des données provenant de plusieurs sources dans un espace de stockage unique comme cela se fait dans un entrepôt de données. Les principaux problèmes qui se posent sont l'intégration des schémas des différentes sources, l'identification des objets, la gestion de la redondance dans les données, et la détection voire la résolution des conflits. Les principales approches d'intégration de données seront décrites dans le chapitre 2.
2. Préparation des données : cela inclut le nettoyage des données, la définition de descripteurs, la réduction de dimension ou la sélection de descripteurs si nécessaire, et le formatage des données en fonction du programme de fouille visé.
3. Fouille de données : l'objectif de cette étape est d'extraire des motifs (ou *patterns*) correspondant à des régularités observées dans les données. Plus généralement, il s'agit de passer des données à des éléments de connaissance réutilisables. Ces éléments de connaissance doivent être nouveaux, non triviaux et non fortuits. Il existe une grande variété de méthodes de fouilles de données que l'on peut classer selon plusieurs critères tels que le type des données fouillées, le type des éléments de connaissance recherchés, le type de techniques utilisées [HK01]. Je décris les principes de quelques méthodes de fouille de données dans la section 1.4.
4. Évaluation et interprétation des résultats de la fouille : cette étape inclut une évaluation voire une validation des résultats de la fouille. Tous les moyens qui peuvent aider l'expert à visualiser ces résultats, à les positionner par rapport au domaine et à aller plus loin dans l'interprétation sont précieux à ce niveau. Ces éléments sont synthétisés au début du chapitre 4.

A chacune de ces étapes et en fonction des résultats, un retour à une des étapes précédentes est possible rendant le processus itératif. La complexité de ce processus requiert une interaction avec l'utilisateur ou analyste. L'analyste est supposé être expert en ECD mais aussi être familier des données et si ce n'est pas le cas, être en interaction étroite avec des experts du domaine. Fayyad *et al.*, ont modélisé le processus d'ECD comme un processus multi-étapes, interactif et itératif [FPSS96] (figure 1.1). Les possibles *instanciations* des différentes étapes de ce modèle de processus sont décrites dans plusieurs ouvrages [HK01, WF05]. Des environnements logiciels couvrant tout ou partie des étapes sont disponibles parmi lesquelles deux plateformes académiques largement utilisées, la plateforme WEKA⁴ [WF05] et la plateforme KNIME⁵ [BCD⁺09].

Par ailleurs, afin que fouille de données ne soit pas synonyme de fouille de fichiers, le processus d'ECD a été modélisé dans le contexte des bases de données par Imielinski et Mannila qui soulignent l'importance de rapprocher le monde des bases de données avec celui de l'extraction de connaissances et introduisent le concept de *bases de données inductives* [IM96]. Nous reviendrons

4. <http://www.cs.waikato.ac.nz/ml/weka>

5. <http://www.knime.org>

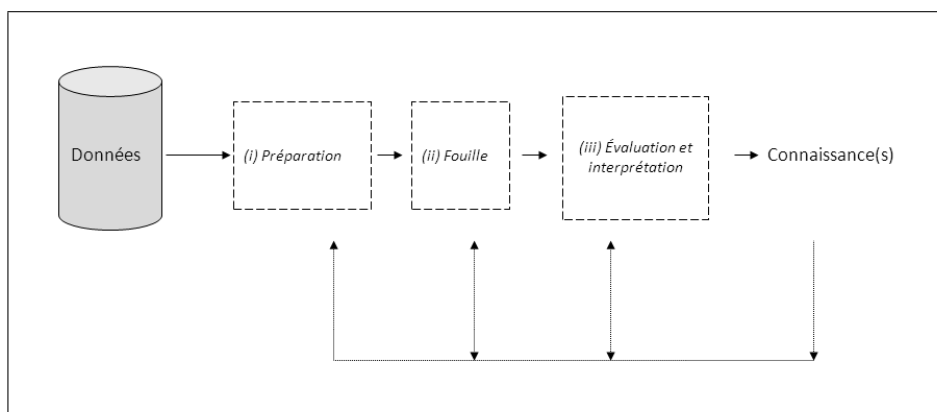


FIGURE 1.1 – Modèle du processus d'ECD inspiré de Fayyad *et al.*, [FPSS96]. L'analyste n'y est pas représenté(e).

sur ce sujet au chapitre 4.

Plus généralement, comme notre bref survol le montrera, le domaine de la fouille de données s'est enrichi de diverses cultures, notamment :

- l'intelligence artificielle qui traite de problèmes complexes (au sens algorithmique) d'apprentissage sur des données de taille modeste (si l'on excepte les travaux récents autour du *deep learning* [Ben09]) ;
- les bases de données qui ont contribué à l'optimisation de certaines tâches de fouille sur des ensembles de données volumineux ;
- les statistiques descriptives que l'on peut (imparfaitement) résumer comme la recherche d'un modèle optimal pour un phénomène, expliquant au mieux les observations de ce phénomène.

1.4 La fouille de données

Différentes méthodes de fouille existent pour différentes tâches d'apprentissage telles la classification supervisée, la classification non supervisée ou clustering, la recherche de régularités (pour décrire ou résumer les données), et la détection de changements ou de déviations (dans des données de séquences). Une méthode de fouille est dite *prédictive* ou *descriptive*. Une méthode *prédictive* travaille sur un ensemble d'exemples classifiés (ou étiquetés) selon un nombre fini de classes. L'objectif de la fouille est alors de produire par induction une hypothèse qui permet de classer correctement les exemples vus mais aussi des nouveaux exemples (on parle de prédire la classe). Une méthode *descriptive* travaille sur des exemples (non nécessairement classifiés) et s'attache à trouver des régularités parmi ces exemples. L'ensemble des régularités peut être vu comme l'hypothèse la plus spécifique qui couvre et explique l'ensemble des exemples [Rae08]. Les méthodes descriptives s'attachent donc à produire des descriptions compactes des données capturant les régularités locales ou globales.

Une distinction peut être faite entre les méthodes dites numériques (incluant la plupart des méthodes statistiques) qui analysent des données numériques (ou quantitatives) et les méthodes symboliques dédiées aux données symboliques (catégorielles ou qualitatives). Évidemment ces deux types de méthodes sont capables de gérer des données quelconques moyennant une conversion des données symboliques en nombres ou une discrétisation des données numériques. Néan-

moins, elles reposent sur des approches et des algorithmes assez différents et produisent des modèles différents. Parmi les exceptions, on trouve la construction d'arbres de décision pour des problèmes de classification supervisée pour laquelle les modèles produits par des algorithmes différents sont similaires (Section 1.4.2).

J'ai privilégié dans mon activité l'utilisation de méthodes symboliques car les données biologiques que nous avons eu à analyser sont principalement symboliques et ces méthodes offrent la possibilité de manipuler des connaissances du domaine (plus ou moins formalisées) mais également de produire des modèles explicites ou auto-explicatifs. Cela n'exclue pas pour autant le recours à des méthodes numériques pour lesquelles il n'y a pas d'équivalent symbolique et qui accomplissent une tâche d'apprentissage importante. C'est le cas des méthodes de classification non supervisée (ou *clustering*) qui opèrent sur des données numériques (Section 1.4.3).

1.4.1 Extraction de motifs et de règles d'association

Mannila et Toivonen ont proposé un cadre qui permet d'englober les principaux travaux dédiés à l'extraction de motifs [MT97]. Nous reprenons ici leurs définitions qui permettront d'introduire les concepts importants dans ce domaine. Etant donné une table de données \mathbf{r} décrivant un ensemble d'objets, un langage L est défini comme un ensemble de motifs. Si I est un ensemble d'items (propriétés booléennes) décrivant des objets, le langage des motifs ensemblistes (ou *itemsets*) L_I comprend tous les sous-ensembles non vides de I tandis que le langage des motifs séquentiels L_S regroupe tous les multi-ensembles possibles de L_I . Ce dernier langage est donc infini contrairement au premier. Une contrainte q est un prédicat booléen défini sur un langage qui définit si un motif m est intéressant ou non. $freq(m) \geq min$ est un exemple de contrainte imposant qu'un nombre minimal de tuples contienne le motif m ($freq(m)$ est la fréquence ou le support du motif m).

Pour un langage L , une table de données r et une contrainte q , la théorie $Th(L, r, q)$ est définie comme l'ensemble des motifs de L qui satisfont la contrainte q dans r .

$$Th(L, r, q) = \{m \in L | q(r, m)\}$$

La recherche efficace de tous les motifs d'une théorie nécessite de structurer l'espace de recherche en définissant un ordre partiel sous forme d'une relation de spécialisation/généralisation entre les motifs. Un motif est plus spécifique qu'un autre motif si tous les objets qui contiennent le premier motif contiennent également le second. Pour les motifs ensemblistes, l'inclusion est une relation de spécialisation (e.g., puisque $X \subseteq XY$, XY est plus spécifique que X).

Une contrainte q respecte la contrainte d'anti-monotonie par rapport à une relation de spécialisation \leq si pour toute table r et tout couple de motifs m et n :

$$\text{si } q(r, m) \text{ et } n \leq m \text{ alors } q(r, n).$$

Les algorithmes d'extraction de motifs par niveaux s'appuient précisément sur l'anti-monotonie de la contrainte par rapport à la relation de spécialisation pour extraire les motifs qui vérifient la contrainte en commençant avec les motifs les plus généraux et en ne considérant pas les spécialisations d'un motif qui ne vérifient pas la contrainte (on parle d'élagage d'une partie de l'espace de recherche).

Mannila et Toivonen décrivent dans ce cadre conceptuel la recherche de motifs ensemblistes et séquentiels. Ils l'appliquent également à la recherche de dépendances fonctionnelles. Si r est une relation comportant les attributs R , X un sous-ensemble de R et A un attribut de R , une dépendance fonctionnelle entre X et A se note :

$$X \rightarrow A$$

Cette dépendance est vraie dans la relation r si pour tout couple de tuples de r ayant la même valeur de X , ils ont la même valeur de A .

Si B et H sont deux motifs alors $B \rightarrow H$ est une règle d'association dont la fréquence est la fréquence du motif HB et dont la confiance est $\frac{freq(HB)}{freq(B)}$ équivalent à $p(H|B)$, la probabilité conditionnelle d'avoir H sachant que l'on a B . Les algorithmes d'extraction de règles d'association requièrent généralement deux paramètres : une fréquence minimale et une confiance minimale. Une fois les motifs fréquents extraits (ceux dont la fréquence est supérieure au seuil), il est possible de les combiner pour produire des règles d'association. L'algorithme APRIORI est le premier algorithme efficace réalisant cette extraction [AS94].

Le nombre de règles extraites peut être très important et des solutions existent pour éviter à l'expert l'analyse de cet ensemble pléthorique de règles. La première solution est une représentation des motifs fréquents sur la base de classes d'équivalence de fréquence [CRB04]. Deux catégories particulières de motifs fréquents sont introduites : les motifs fermés (un motif est fermé si toutes ses spécialisations strictes ont une fréquence strictement inférieure) et les motifs libres (un motif est libre si toutes ses généralisations strictes ont une fréquence strictement supérieure). Ces représentations condensées permettent de réduire l'ensemble des motifs fréquents extraits mais aussi de faciliter l'extraction des règles d'association. Par exemple, les motifs fermés et libres permettent l'extraction de règles d'association particulières (prémisse minimale et conclusion maximale) [BPT⁺00]. Des contraintes exprimant mieux les besoins des utilisateurs (agrégations, modèles de règles. . .) ont été étudiées ainsi que les possibilités de les combiner et de les *pousser* dans le processus d'extraction de motifs ou de règles d'association [NLHP98, PHL04, BM05, Sou06].

D'autres mesures que la fréquence et la confiance ont été définies afin de classer les règles d'association telles que le *lift* [BMS97]. La mesure de *lift* (le *lift* de $B \rightarrow H$ est défini comme $\frac{freq(HB)}{freq(H)freq(B)}$) permet de mieux quantifier la corrélation statistique entre deux motifs. Une littérature abondante existe sur le sujet, notamment l'étude comparative réalisée par Tan *et al.*, des différentes mesures d'intérêt proposées en statistique, en fouille de données, et en apprentissage [TKS02]. Une évaluation rigoureuse de la significativité statistique d'un résultat de fouille est primordiale notamment pour les applications bio-médicales. En effet, une proportion variable de motifs ou de règles est trouvée par hasard selon la densité du jeu de données analysé et les paramètres de l'algorithme de fouille [FPSS96]. Des travaux ont ainsi porté sur la façon d'adapter la méthodologie des tests statistiques à la fouille de données et précisément à la fouille de motifs ensemblistes et de règles d'association. Ces travaux reposent sur un échantillonnage des données ou sur une randomisation spécifique des données et sur une correction permettant de contrôler le test d'hypothèses multiples et de limiter le risque de considérer à tort au moins un des motifs ou une des règles comme significatif [GMMT07, LTP07, Web07].

Extraire des règles d'association sans *a priori* permet de réaliser une *induction descriptive* puisqu'il s'agit de trouver des régularités qui *résumant* les données. Si l'on se limite à l'extraction

de règles dont la partie droite comporte un attribut d'intérêt, nous basculons vers une *induction prédictive* et les règles recherchées sont appelées *règles de classification*, même si une règle d'association ne révèle pas nécessairement un lien de causalité. C'est une façon de superviser la recherche de règles en étiquetant les objets par une information de classe. Cela a donné lieu à divers travaux portant sur la recherche de motifs discriminants, de motifs émergents, et de sous-groupes. Un cadre unifiant l'ensemble de ces travaux comme une recherche supervisée de règles descriptives a été proposé [NLW09]. De même, diverses mesures d'évaluation de règles qu'elles soient descriptives ou prédictives (confiance, précision, sensibilité, spécificité, nouveauté...) avaient été unifiées par Lavrac *et al.*, [LFZ99]. La mesure pondérée de précision relative *WRAcc* a été proposée à cette occasion comme un compromis entre précision, sensibilité et nouveauté d'une règle. La mesure *WRAcc* d'une règle $B \rightarrow H$ est définie comme :

$$WRAcc(B \rightarrow H) = p(B)(p(H|B) - p(H)) = freq(HB) - freq(H)freq(B)$$

1.4.2 Classification supervisée

Étant donné un ensemble d'objets (ou exemples) décrits par un ensemble de propriétés et étiquetés selon leur appartenance à une classe, le but de la classification supervisée est la construction d'un modèle permettant de prédire l'information de la classe à partir des valeurs de certaines propriétés et ce, aussi bien pour les exemples d'apprentissage que pour de nouvelles observations. On parle de classification binaire lorsque deux classes sont considérées et de régression lorsque la valeur à prédire est une grandeur continue.

L'algorithme le plus simple de classification ou de régression est certainement celui des *k*-plus proches voisins (*KNN*). La règle de classification consiste simplement à prédire la valeur de la classe d'un exemple en utilisant la valeur attachée aux *k* objets les plus proches de l'exemple en utilisant les distances entre paires d'exemples. La prédiction se fait selon un vote majoritaire (valeur de la classe la plus fréquente) en cas de classification ou une moyenne pour la régression. Le plus difficile est le choix de la valeur de *k* et de la distance.

Une autre méthode de classification supervisée produisant un modèle explicite est la construction d'un arbre de décision selon l'algorithme proposé par Quinlan [Qui96]. Parallèlement, Breiman *et al.*, proposaient une méthode de construction d'arbres de classification et de régression fondée sur une approche similaire aux arbres de décision sur des données numériques [BFOS84]. Les données sont représentées sous forme d'une matrice d'objets dans laquelle chaque objet est décrit par une série d'attribut-valeurs et est étiqueté par le label de sa classe. L'algorithme qui permet la construction de l'arbre de décision est de type *diviser pour régner* et consiste à chaque étape à choisir (selon une certaine méthode), parmi une liste d'attributs, l'attribut qui permet de discriminer le mieux un ensemble d'objets par rapport à leur classe. Des arbres compacts sont préférables comme outil d'aide à la décision et des mécanismes d'élagage permettent de simplifier l'arbre obtenu et de réduire le sur-ajustement par rapport aux données d'apprentissage (*overfitting*) [WF05]. Lorsque le nombre d'attributs est important, la méthode (d'ensemble) des forêts aléatoires (*random forests*) consiste à tirer au hasard de nombreux sous-ensembles d'attributs sur lesquels des arbres de décision sont construits [Bre01]. La prédiction de la classe pour une nouvelle instance se fait alors par un vote sur la base des prédictions des différents arbres.

Comme nous l'avons mentionné dans la section précédente, il est possible de produire des règles de classification par les algorithmes d'extraction de règles d'association. Ces règles consti-

tuent des modèles locaux qui ne couvrent souvent qu'une partie de l'espace de données contrairement à un arbre de décision qui constitue un modèle global couvrant l'ensemble des données.

L'apprentissage d'un concept permet également de généraliser les données d'observation relatives à un ensemble d'exemples (instances du concept assimilé à une classe) et d'exemples négatifs (qui ne sont pas instances du concept) [Mit82]. La généralisation consiste à trouver, dans le cadre d'un langage de représentation donné, une hypothèse qui couvre tous les exemples positifs et ne couvre pas d'exemple négatif. Des formalisations de l'apprentissage de concept ont été proposées en logique des prédicats ou sur la base du modèle relationnel de données et nous y reviendrons dans le chapitre 3.

1.4.3 Classification non supervisée (clustering)

La classification non supervisée consiste, étant donné un ensemble d'objets décrits selon un ensemble de propriétés (ou variables), à grouper ces objets par similarité. Ainsi, des groupes (ou clusters) sont construits qui regroupent les objets proches et séparent les objets éloignés. Généralement, le nombre de clusters est une donnée en entrée de l'algorithme de classification.

Les algorithmes de clustering se distinguent tout d'abord par la nature des données sur lesquelles ils opèrent. En effet, certains algorithmes travaillent sur les vecteurs de caractéristiques décrivant les objets à grouper. L'algorithme est alors en mesure de calculer des distances (de type L^p) ou des (dis)similarités entre objets pour réaliser le groupement. La nature des variables (continues à échelle d'intervalle ou de rapport, binaires, nominales, ordinales) devrait influencer sur la manière de définir la mesure de distance ou de (dis)similarité entre deux objets. Kaufman et Rousseeuw proposent une mesure générique de dissimilarité tenant compte de la nature de chaque variable [KR05]. D'autres algorithmes peuvent opérer directement sur une matrice carrée de (dis)similarités entre objets.

Si nous nous limitons aux algorithmes de base pour le clustering, nous pouvons en distinguer deux types, les algorithmes par partitionnement et les algorithmes hiérarchiques [KR05].

Étant donné un nombre de clusters, les méthodes par partitionnement affectent chaque objet à un cluster en minimisant sa distance au centroïde du cluster (K-means) ou au médoïde du cluster. Le médoïde est un des objets à grouper dont la localisation dans le cluster est centrale (minimise la distance aux objets du cluster). À l'issue du partitionnement, on dispose donc d'un représentant pour chacun des k clusters construits. Le centroïde est, quant à lui, un objet virtuel calculé comme la moyenne des vecteurs des objets du cluster. Par conséquent, le partitionnement autour de médoïdes peut fonctionner sur une matrice de dissimilarité tandis que le partitionnement autour de centroïdes nécessite de disposer des vecteurs caractéristiques nécessaires pour la construction des centroïdes et pour le calcul des distances entre objet et centroïde. Lorsque l'on suppose que les k clusters à trouver sont chevauchants, des algorithmes spécifiques ont été proposés pour calculer des coefficients d'appartenance de chaque objet à chaque cluster tels que le *C-means* et l'algorithme FANNY capable de traiter des matrices de dissimilarité [KR05]. Lorsque la forme des clusters à rechercher n'est pas convexe, des méthodes à base de la notion de densité ont été proposées telles que l'algorithme DBSCAN [EKX96]. Dans ce cas, les clusters correspondent à des régions denses dans un espace à n dimensions séparées par des régions de moindre densité, ces dernières correspondant au bruit dans les données.

Les méthodes hiérarchiques agglomératives ou ascendantes (HAC) opèrent sur une matrice de dissimilarité et construisent à chaque étape des clusters en fusionnant les clusters les plus proches trouvés à l'étape précédente (on commence avec n clusters correspondant à chacun des n objets à grouper). Différentes versions ont été proposées selon la définition de la proximité entre deux clusters. Il est à noter que les clusters obtenus peuvent ne pas être optimaux car les décisions prises à une étape ne peuvent plus être remises en cause aux étapes suivantes. Néanmoins les biologistes apprécient les méthodes hiérarchiques et les ont beaucoup exploitées pour la construction de taxonomies. Elles offrent le double avantage de ne pas nécessiter de fixer un nombre de clusters et de produire un résultat assez visuel sous forme de dendrogrammes. En dehors de la construction de taxonomies, une méthode HAC peut toutefois servir à se faire une idée du nombre de clusters que l'on recherche ensuite avec des méthodes de partitionnement. Une utilisation très répandue de ce type de clustering associée à une visualisation sous forme de carte de chaleur (*heatmap*) est celle qui permet l'analyse des résultats d'expérience de transcriptomique à l'aide de puces à ADN [ESBB98]. La littérature sur le sujet de la classification non supervisée est extrêmement abondante et de nombreuses versions des algorithmes de base ont été proposées [ELLS11].

Lorsque l'on souhaite grouper des objets biologiques complexes, il est possible de rendre compte de cette complexité grâce à une mesure de similarité pouvant prendre en compte la connaissance de domaine. Les méthodes de clustering capables d'opérer sur la matrice de similarité (telles que le partitionnement autour de médoides) sont alors indiquées puisqu'elles n'obligent pas à représenter les objets sous forme de vecteurs caractéristiques dans lesquels les caractéristiques sont supposées numériques et indépendantes. Les noyaux (*kernels*) offrent la possibilité de définir des mesures de similarité sur des données structurées (telles que des vecteurs, des multi-ensembles, des arbres, des graphes) et peuvent s'avérer utiles pour des données biologiques [SS02, BOS⁺05, GTDdB07].

1.4.4 Analyse formelle de concepts

L'analyse formelle de concepts (*Formal Concept Analysis* ou *FCA*) est décrite dans [GW99]. La FCA s'applique à un contexte formel défini comme un triplet (G, M, I) où G dénote un ensemble d'objets, M un ensemble d'attributs et $I \subseteq G \times M$ une relation binaire entre G et M . $(g, m) \in I$ est interprété comme "l'objet g possède l'attribut m " (noté aussi gIm). Deux opérateurs $(.)'$ définissent une connexion de Galois entre les ensembles de parties d'ensembles $(2^G, \subseteq)$ et $(2^M, \subseteq)$, avec $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : gIm\}$$

$$B' = \{g \in G \mid \forall m \in B : gIm\}$$

Pour $A \subseteq G, B \subseteq M$, la paire (A, B) , telle que $A' = B$ et $B' = A$, est un *concept formel*. L'ensemble A est l'*extension* et l'ensemble B est l'*intension* du concept (A, B) . L'ensemble des concepts formels est partiellement ordonné par la relation de subsomption de concept :

$$(A1, B1) \leq (A2, B2) \Leftrightarrow A1 \subseteq A2 \Leftrightarrow B2 \subseteq B1$$

L'ensemble des concepts formels muni de cette relation d'ordre partiel forme un treillis appelé treillis de concept (ou treillis de Galois) associé au contexte (G, M, I) . La table 1.2 contient un exemple de contexte formel très simple (tiré de [10]) et la figure 1.2 représente le treillis de

	rule_1	rule_2	rule_3
patch_1	x		x
patch_2			x
patch_3	x	x	
patch_4	x	x	x
patch_5		x	
patch_6	x		x
patch_7		x	x
patch_8			

TABLE 1.2 – Exemple de contexte formel.

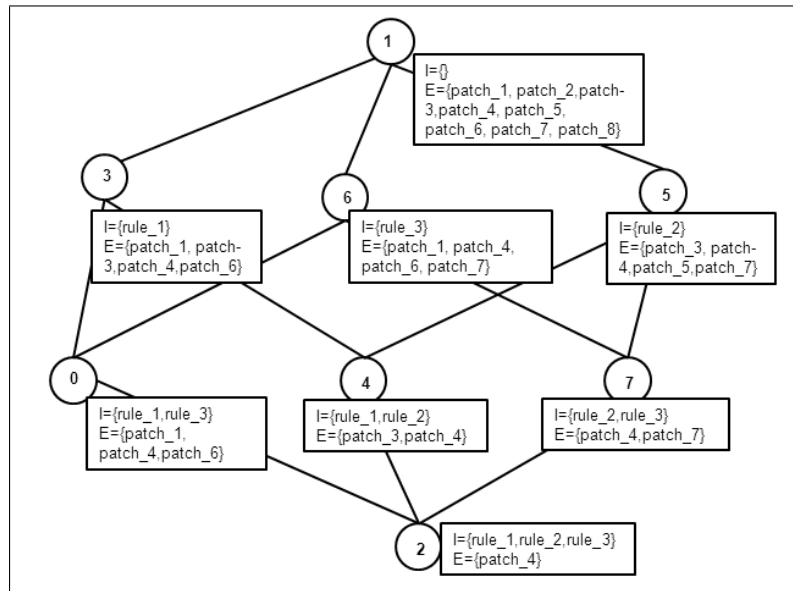


FIGURE 1.2 – Le treillis de concepts obtenu par FCA sur la table 1.2. Chaque concept numéroté a une intension (I) et une extension (E).

concepts construit à partir de ce contexte formel.

La FCA est une théorie mathématique qui a été exploitée notamment pour l'analyse de données, l'extraction de connaissances, l'acquisition de connaissance. De nombreux travaux fondamentaux sont consacrés à diverses extensions de la théorie initiale pour la prise en compte de matrices de données quelconques (*e.g.*, *pattern structures*, analyse triadique). Kuznetsov et Obiedkov passent en revue les algorithmes qui ont été développés pour la construction de treillis de concepts [KO02]. L'implémentation de la FCA que nous utilisons est celle de la plateforme Coron (<http://coron.loria.fr>) développée par Laszlo Szathlmary [Sza06].

En pratique, ce type d'approche produit des treillis volumineux lorsque les données sont volumineuses et denses. Mais lorsque les données sont peu volumineuses et *binaires de façon native*, l'AFC est une méthode puissante qui permet, par exemple, d'amorcer ou d'enrichir la construction d'une ontologie de domaine à partir de données, sous la forme d'une hiérarchie de classes avec une définition de chaque classe en intension et en extension [CHST04, BTN08]. La FCA est également précieuse pour l'interprétation ou le raffinement de résultats de fouille obtenus par d'autres méthodes, ce qui est parfois appelé analyse secondaire (j'y reviendrai dans le chapitre 4).

1.5 Plan du document

Le reste du document est organisé en quatre chapitres. Le chapitre 2 porte sur l'intégration des données dirigée par un modèle et la préparation de ces données en amont de la fouille. Le chapitre 3 concerne la fouille de données relationnelles et la façon dont nous l'avons appliquée à différents problèmes biologiques. Le chapitre 4 porte sur l'évaluation et l'interprétation des modèles résultant de la fouille ainsi que les possibilités d'enchaînement avec une itération nouvelle du processus d'ECD. Chaque chapitre débutera par un bref état de l'art et sera suivi de notre contribution. Enfin, le chapitre 5 conclut ce document par un projet de recherche sous forme d'un ensemble de perspectives de recherche à court et à long terme.

Chapitre 2

Préparation pour la fouille des données biologiques

2.1 Introduction

Intégrer des données revient à combiner les données de plusieurs sources hétérogènes dans un espace cohérent. L'objectif de l'intégration est soit l'accès simplifié aux différentes sources soit la possibilité d'effectuer des analyses sur les données intégrées. Les principales approches d'intégration de données ainsi que leur application sur les données biologiques seront décrites dans la section 2.3. Une étape préalable à l'intégration est l'identification des sources de données susceptibles d'être intéressantes pour l'objectif fixé. Cela est connu sous le vocable de découverte de ressources. La section 2.2 présente de façon synthétique les efforts de structuration des ressources biologiques afin de faciliter leur recherche. Notre contribution par rapport à ces aspects du processus d'ECD est résumée dans la section 2.4.

2.2 Organisation du ressourceome biologique et découverte de sources de données

Le concept de *ressourceome* (par analogie avec génome, protéome, métabolome) a été introduit par Cannata et al. pour désigner les programmes et les sources de données biologiques disponibles sur le web et insister sur l'importance d'annoter et d'organiser ces ressources afin d'en optimiser l'accès et l'utilisation [CMA05].

La découverte de sources de données intéressantes peut difficilement se faire à l'aide des moteurs de recherche fondés sur le contenu. Ces moteurs n'étant pas capables de distinguer les pages servant d'interface à des sources de données des documents mentionnant ces sources, se révèlent inefficaces pour cette tâche. Des *crawlers* spécifiques ont été proposés afin d'identifier les formulaires d'interrogation et de typer les sources de données selon le contenu textuel des formulaires [KSS⁺07, BF06]. L'indexation qui en résulte reste néanmoins trop pauvre pour permettre une recherche efficace.

Plusieurs catalogues de ressources biologiques existent sur le web dont les plus rudimentaires sont des portails. Un portail est une compilation plus ou moins structurée de liens vers des serveurs web donnant accès à des bases de données ou à des programmes. Généralement

maintenus par un individu, ce type de portails malgré leur intérêt, restent difficiles à exploiter par des personnes sans connaissance a priori sur les ressources. L'exhaustivité et le maintien à jour de ces portails n'étant pas garantis, leur qualité est susceptible de se dégrader rapidement. Quelques catalogues documentés et mieux maintenus existent grâce à des efforts collaboratifs tels que *Bioinformatics Links Directory*, le catalogue de la revue NAR (mentionné dans le chapitre 1) [BYYO11]. Néanmoins, l'interrogation de ces catalogues se limite le plus souvent à une recherche textuelle dans le titre, la description ou l'URL des ressources.

Afin d'optimiser la recherche ou la découverte de ressources, des méta-données pertinentes doivent être rassemblées dans des annuaires selon les principaux critères de description des ressources. Une méta-donnée est une donnée servant à définir ou à décrire une autre donnée. Les méta-données servent depuis plusieurs siècles à codifier le signalement et le contenu de documents afin de faciliter leur recherche. Les premiers catalogues de méta-données sur les ressources biologiques (BioCat, DBCAT, BioNetBook...) ont eu pour objectif de répertorier les programmes d'analyse ou les bases de données. La limite de ces catalogues réside dans le coût de la maintenance (qui a conduit à leur abandon) et dans la définition des vocabulaires utilisés pour la représentation des champs de méta-données. En effet, un vocabulaire ouvert ou un vocabulaire fermé mais trop restreint dégrade les performances de la recherche. En 1995, la standardisation des méta-données pour décrire les ressources du web fait l'objet d'un groupe de travail à Dublin (Ohio, USA) d'où est issu le standard DCMI (Dublin Core Metadata Initiative) comportant une quinzaine d'éléments de description d'une ressource (auteur, titre, date de création, contenu, adresse, etc.) [DW03]. En 1999, le W3C (World Wide Web Consortium) propose le langage RDF (Resource Description Framework), décrit selon la syntaxe XML, permettant d'attacher des méta-données à des ressources du web.

Un nombre croissant de ressources biologiques, incluant programmes d'analyse et accès à des bases de données, deviennent accessibles selon des interfaces d'invocation standard (interfaces de programmation) sous forme de services web. Deux projets BioMoby et MyGrid à l'initiative respectivement des universités de Saskatchewan (Canada) et de Manchester (Grande Bretagne) facilitent l'accès à ces ressources et l'automatisation des procédures d'analyse ou d'interrogation de bases de données impliquant plusieurs étapes stéréotypées sous forme de workflows [Con08, GWS03]. Ces plateformes sont conformes à l'architecture orientée service (*Service-Oriented Architecture* ou SOA) très utilisée dans les applications commerciales et reposant sur (i) un annuaire de méta-données décrivant les services, (ii) un outil de recherche de services, et (iii) une procédure standard d'invocation de ces services. A la fin des années 2000, ces plateformes réunies donnent accès à près de 5000 services web biologiques. Le projet MyGrid a permis de développer un module de composition et d'exécution de workflows à l'image des pipelines d'expériences biologiques. Les entrées et les sorties des services web y sont décrits à l'aide d'ontologies de domaine construites de façon ad hoc et exprimées en RDFS. La recherche de services répondant à une requête est alors vue comme un problème d'appariement de méta-données sémantiques et laisse au biologiste la responsabilité du choix final des services pertinents parmi les services retrouvés [LAWG05].

L'état de l'art nous a poussés à proposer une structure d'annuaire en tentant de dépasser les limites des annuaires existants pour faciliter la tâche du biologiste dans sa quête de sources de données que son objectif soit d'interroger individuellement chaque source identifiée ou d'intégrer les données provenant de plusieurs sources identifiées (Section 2.4.1).

2.3 Intégration des données biologiques

Deux approches existent pour l'intégration de données selon la localisation des données :

- *l'approche matérialisée* pour laquelle les données sont dans un entrepôt de données où elles sont rapatriées depuis leur source d'origine ;
- *l'approche virtuelle* pour laquelle les données restent dans les sources distribuées et elles sont interrogées ou accédées par le biais d'un système médiateur.

L'approche matérialisée ou entrepôt de données consiste en la construction d'une base de données appelée entrepôt pour stocker les données provenant de différentes sources. Cette solution est très utilisée dans le commerce et la gestion où les entrepôts servent de support d'aide à la décision grâce aux analyses de type OLAP (*On-Line Analytical Processing*) qu'ils permettent par opposition aux analyses de type transactionnel (OLTP, *On-Line Transactional Processing*) que les SGBD offrent [AAD⁺96]. Dans une approche entrepôt de données, l'intégration s'appuie sur un schéma de données multi-dimensionnel défini spécifiquement pour l'entrepôt. Les données sont alors extraites des sources, transformées et nettoyées avant d'être chargées dans l'entrepôt (on parle de processus ETL pour *Extract, Transform, Load*). L'utilisateur peut interroger directement l'entrepôt ou interagir avec l'entrepôt par l'intermédiaire de vues définies sur les données [BCT06]. Ces vues s'appuient sur un ensemble prédéfini d'opérateurs d'agrégation et permettent de visualiser les données sous forme de cubes [AAD⁺96]. L'entrepôt Gedaw est un exemple de mise en oeuvre de l'approche entrepôt pour l'intégration et l'analyse de données du transcriptome humain [GMB⁺05]. Les systèmes BioMart et BioWarehouse sont des systèmes plus généraux d'intégration de données biologiques suivant une approche entrepôt [KKS⁺04, KLW08].

En ce qui concerne l'approche médiateur, l'intégration de données est fondée sur la définition d'un schéma global unifiant les schémas hétérogènes des sources à intégrer [Wie92]. La description d'un tel schéma implique la mise au point de correspondances ou *mappings* souvent conceptualisés comme des vues au sens des bases de données relationnelles (une vue est une requête faisant appel aux opérateurs de l'algèbre relationnelle appliquées à un ensemble de schémas de relations) [Len02]. L'approche *Local as View* consiste à définir le contenu de chaque source locale de données comme une vue sur le schéma global tandis que l'approche *global as View* définit le schéma global comme une vue sur les schémas des sources intégrées. La tâche du médiateur consiste à calculer les réponses à une requête exprimée sur le schéma global par ré-écriture de cette requête en utilisant les vues. Il s'agit de trouver une requête équivalente à la requête initiale ou qui implique logiquement cette requête puis évaluer cette requête ré-écrite sur les extensions courantes des vues [Ull00, RBF⁺02].

Selon Lenzerini, le web idéal serait donc composé de systèmes de médiation intelligents capables de prendre en charge une requête utilisateur, la distribuer à plusieurs bases de données pertinentes, collecter les réponses partielles et les fusionner avant de les proposer à l'utilisateur [Len02]. Les principes des systèmes de médiation ont été appliqués aux données biologiques et cela a donné lieu à divers systèmes tels que TAMBIS, BIS ou BioMediator [SBB⁺00, LBE03, MSTH05]. Cette liste n'est pas exhaustive et démontre bien le caractère insatisfaisant des systèmes produits. Ainsi aucun de ces systèmes n'a suscité un réel engouement de la part des biologistes probablement à cause de la sophistication de la mise en oeuvre de ces systèmes mais surtout du nombre limité de sources de données interfacées.

L'approche entrepôt se distingue de l'approche médiation par le fait que la première a pour objectif de réunir physiquement les données en vue d'analyses diverses tandis que la seconde a

pour objectif d'offrir un accès unifié à des sources hétérogènes en donnant l'illusion à l'utilisateur qu'il interagit avec une source unique. Les biologistes ont besoin de médiateurs afin de tenter de maîtriser le foisonnement des sources de données disponibles [BBDF07]. Pour ma part, je m'intéresse à l'intégration de données dans le contexte du processus découverte de connaissances et c'est donc vers l'approche entrepôt que je me suis tournée de préférence (Section 2.4.2)

2.4 Contribution

Afin d'assurer une bonne couverture des sources de données et de répondre aux mieux aux attentes des biologistes, nous considérons que le problème de découverte de sources de données doit être traité et optimisé indépendamment de celui de l'intégration ou de la conception de workflows de collecte de données. Pour cela, nous avons proposé d'organiser et de faciliter la découverte de bases de données biologiques grâce à un annuaire, BioRegistry (Section 2.4.1). Afin de faciliter ensuite, en amont de la fouille, la collecte et l'intégration de données à partir de plusieurs BDs, nous avons proposé une approche qui prend le modèle de données comme un paramètre en entrée (Section 2.4.2). La dernière contribution est la définition d'une mesure de similarité entre objets biologiques et son utilisation de deux façons différentes au service de cette préparation des données (Section 2.4.3).

2.4.1 BioRegistry : Organisation et découverte de bases de données biologiques

BioRegistry : un annuaire de sources de données biologiques Nous avons conçu et mis en œuvre un modèle d'annuaire de bases de données biologiques appelé BioRegistry [63] dont l'objectif était de dépasser les limites des catalogues précédents grâce à certaines caractéristiques. Tout d'abord, le modèle de méta-données est une adaptation du standard DCMI pour la description des bases de données biologiques, incluant leur identification, leur contenu, et leur qualité. Des ontologies de domaine sont utilisées pour l'encodage des méta-données. Ensuite, l'annuaire est construit en partant d'un catalogue reconnu de bases de données biologiques, à savoir le catalogue NAR mentionné ci-dessus (et présenté dans la section 1.2.1). Nous ajoutons de la valeur à ce catalogue en enrichissant l'indexation factuelle des bases de données avec des méta-données exprimées dans le vocabulaire idoine. Nous avons choisi de décrire le contenu d'une base de données biologique à l'aide du thésaurus MeSH utilisé pour décrire le contenu des publications bio-médicales. Le thésaurus MeSH est maintenu par la NLM et comporte près de 23000 descripteurs structurés en 15 catégories. La plupart de ces catégories sont pertinentes pour caractériser le contenu d'une source de données : "Organisms" [B], "Diseases" [C], "Chemicals and Drugs" [D], "Biological Sciences" [G], "Natural Sciences"[H], etc. Quelques catégories ne le sont pas, telles que "Information Science" [L], "Geographic Locations"[Z], "Humanities"[K]. Le thésaurus MeSH permet de structurer les termes de chaque catégorie à l'aide de relations de généralisation/spécialisation et de capturer en partie la connaissance biologique.

La figure 2.1 présente le modèle de la base de données qui structure et héberge les données de l'annuaire BioRegistry.

Nous avons anticipé le problème de la maintenance de l'annuaire grâce à des procédures automatiques d'extraction des méta-données lors des mises à jour de la sources primaire d'information, à savoir le catalogue de la revue NAR (mis à jour annuelle). Ainsi une liste de termes MeSH sont extraits de notices bibliographiques décrivant la source et à partir de l'analyse de la description textuelle de la source de données [25, 24]. Ce travail a été fait en partie dans le cadre

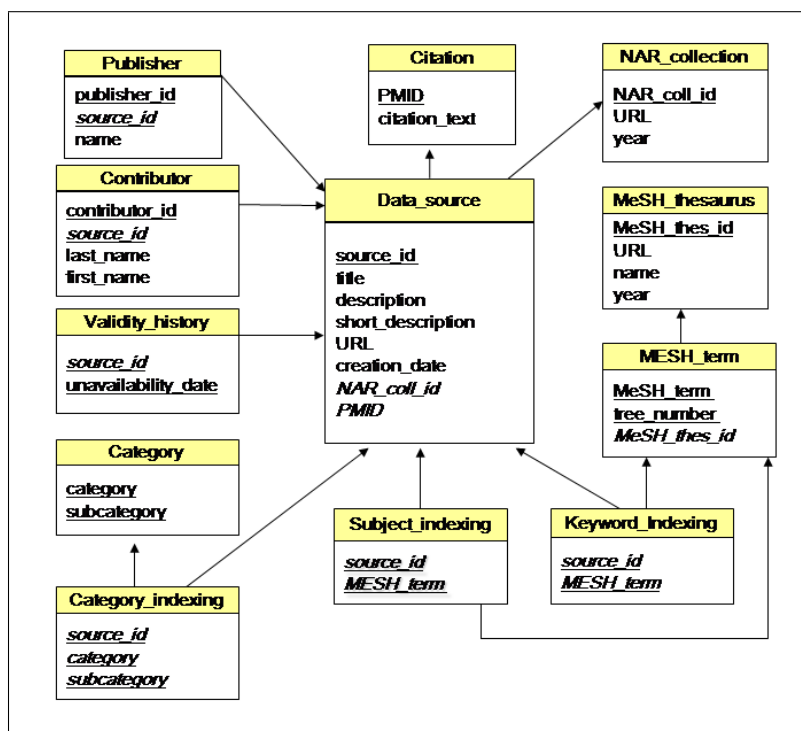


FIGURE 2.1 – Modèle de données de l'annuaire BioRegistry (figure tirée de [24]).

du mémoire CNAM de Philippe Franiatte que j'ai encadré (soutenu en 2009).

Une façon d'exploiter l'annuaire BioRegistry pour la recherche de sources de données biologiques est de permettre une recherche par requête incluant une mesure de similarité sémantique permettant d'exploiter la connaissance du domaine encodée dans le thésaurus MeSH. De nombreuses mesures de similarité ont été proposées tenant compte de relations hiérarchiques entre termes d'indexation. Martin *et al.*, ont notamment défini une mesure de similarité entre deux listes de termes de Gene Ontology comptabilisant le pourcentage de termes communs entre les deux listes enrichies préalablement de tous les termes ancêtres de chaque terme de chaque liste [MBR⁺04]. Nous avons adapté cette mesure pour comparer la liste de termes MeSH formant la requête de l'utilisateur avec la liste de termes qui décrit chaque base de données dans l'annuaire.

Classification et découverte de sources de données biologiques Nous avons déjà regretté l'absence d'une classification objective des bases de données biologiques. Idéalement une classification flexible et automatique permettrait à l'utilisateur de découvrir la (les) base(s) de données pertinente(s) par un processus de navigation. De plus, la recherche de sources exclusivement par requête est pertinente lorsque le besoin est assez bien défini. Dans le cas contraire, il peut être plus aisé pour le biologiste de reconnaître une BD pertinente que de la décrire. Cela a motivé, dans le cadre de la thèse de Nizar Messai, l'utilisation et l'extension de l'analyse formelle de concepts pour la classification automatique des bases de données contenues dans l'annuaire BioRegistry et l'aide à la découverte par requête ou par navigation dans le treillis de concepts obtenu [43, 54, 44, 46, 53, 50, 52, 51].

Tout d'abord, un treillis de concepts est construit sur la base des méta-données de l'annuaire

transformées en contexte formel (chaque source de donnée est décrite par un ensemble des critères booléens correspondant aux divers termes MeSH). Ce treillis constitue une classification des sources de l'annuaire et peut se prêter à une navigation exploratoire si sa taille n'est pas prohibitive.

Un algorithme complet et correct BR-Explorer (BioRegistry Explorer) a été défini pour la recherche de sources pertinentes par rapport à une requête-utilisateur et consiste à :

1. insérer une requête utilisateur (simple ensemble de mots-clés correspondant aux critères de recherche) comme un nouvel objet dans le treillis. Pour ce faire, le treillis est construit puis mis à jour à l'aide d'un algorithme incrémental de construction de treillis de concepts ;
2. localiser le concept pivot (celui dont l'intension est la requête) dans le treillis
3. classer les extensions du concept pivot et de ses super-concepts par rapport au nombre de critères partagés avec la requête

Divers mécanismes de prise en compte de la connaissance de domaine ont été proposés pour enrichir l'algorithme BR-Explorer. Ainsi, le raffinement de requête par généralisation ou par spécialisation des critères de la requête en faisant appel à des ontologies de domaine (MeSH, Taxonomie des organismes vivants) permet de proposer des sources en cas d'échec de l'algorithme BR-Explorer [43, 44, 45]. Afin d'améliorer l'expressivité dans la requête-utilisateur, il est possible d'exprimer des relations hiérarchiques entre attributs pour exprimer l'importance relative des critères de sa requête (par exemple, les critères décrivant le contenu d'une source peuvent être déclarés comme plus importants que ceux qui décrivent sa qualité). La prise en compte d'une telle hiérarchie d'attributs dans un treillis de concepts consiste à restreindre la recherche dans le treillis aux concepts qui vérifient les dépendances exprimées entre attributs. Un concept vérifie une dépendance entre attributs lorsque son intension ne contient pas un attribut secondaire sans l'attribut principal duquel il est dépendant. Un tel concept est dit cohérent vis-à-vis de la dépendance considérée [49, 53].

Une extension plus majeure a porté sur la gestion de contextes de données multi-valués. En effet, une façon naturelle de décrire les sources de l'annuaire est de considérer chaque catégorie du MeSH présente dans l'annuaire (*Organisms* [B], *Diseases* [C], *Chemicals and Drugs* [D]...) comme un attribut et d'associer à chaque source l'ensemble des termes MeSH qui la décrivent selon cette catégorie. L'algorithme SimBA (Similarity-Based Complex Data Analysis System) permet de construire un treillis multi-valué de concepts en utilisant une mesure de similarité entre les valeurs d'attributs [50]. Selon cette nouvelle définition de la connexion de galois, deux objets partagent une propriété si les valeurs de cette propriété pour ces objets sont similaires et plus précisément si cette similarité est supérieure à un seuil fixé par l'utilisateur.

L'algorithme SimBA a été appliqué au contexte multi-valué issu de l'annuaire en utilisant la mesure de similarité sémantique mentionnée dans la section précédente avec différents seuils de similarité. Des opérations de Zoom avant/arrière permettent alors d'explorer progressivement ces treillis qui constituent une classification conceptuelle (à précision variable) des sources de données biologiques [52]⁶.

6. Publication jointe à ce document

2.4.2 MODIM : Intégration de données fondée sur un modèle en vue de la fouille de données

Une fois identifiées les sources de données pertinentes pour un problème d'ECD, il faut s'atteler à l'intégration de données provenant de ces sources. Nous avons pu constater que les biologistes ont tendance à utiliser des tableurs pour structurer leurs données en classeurs pour gérer les tableaux de données issues d'un mélange d'étapes manuelles d'exploration et de scripts d'analyse. Il est évident, de notre point de vue d'informaticien, qu'il faut préférer un système de gestion de données plus évolué que le tableur. Les SGBD (Systèmes de Gestion de Bases de Données) relationnels, grâce à leurs fonctions de définition (schémas de relations incluant un ensemble de contraintes d'intégrité) et de manipulation de données (interrogation structurée, agrégations, mise à jour), sont une excellente alternative. Une autre fonction des SGBD permet de garantir l'intégrité des données vis-à-vis des contraintes exprimées (contraintes d'unicité, de référence...). Cette alternative des SGBD est d'autant plus intéressante que l'offre logicielle est très riche y compris dans le monde du logiciel libre et que les SGBD actuels offrent de très bonnes performances pour la gestion de gros volumes de données.

Comme pour les entrepôts, l'intégration des données dans le contexte d'un processus d'ECD a pour objectif de préparer les données à l'analyse et pas seulement pour l'interrogation et l'accès. Néanmoins, pour notre problème d'intégration de données biologiques comme étape d'un processus d'ECD, nous n'avons pas opté pour une intégration sous forme d'entrepôt en utilisant les logiciels existants pour trois raisons :

- La préparation des données intégrées à partir de bases de données s'avère assez difficile à cause de la grande hétérogénéité qui y règne et de l'absence d'un niveau sémantique dans les modèles de données. Clairement le processus ETL préconisé dans les systèmes d'entrepôt n'est pas adapté (les ETL clés-en main sont trop restrictifs et les ETL développés à façon sont trop ouverts).
- Les entrepôts privilégient des analyses de données multi-dimensionnelles à l'aide d'opérateurs d'agrégation s'appliquant à des données comportant des attributs hiérarchiques (temps, produit, adresse). Dans le cas des données biologiques, le modèle de données est difficile à contraindre a priori et le type d'analyse à effectuer doit rester ouvert.
- La spécification du processus d'intégration devrait être dirigée par le modèle de données et privilégier un mode déclaratif (les programmes de collecte devraient être générés automatiquement).

Nous avons donc défini une approche appelée MODIM (Model-driven Data Integration for Mining) comme une solution pragmatique d'intégration de données dans un contexte d'ECD. Nous avons privilégié une approche que l'on peut qualifier d'*agile*⁷ pour l'intégration afin de répondre aux besoins (parfois fluctuants) des biologistes du fait que les sources de données primaires sont instables en termes de modèles d'exposition des données mais aussi en termes de sources disponibles (apparition régulière de nouvelles ressources et disparition de certaines ressources, faute de maintenance).

Principes de l'approche MODIM : Le logiciel MODIM met en œuvre une approche pour collecter et intégrer des données en fonction d'un modèle relationnel défini au préalable [55]. Ce modèle de données est construit (par les biologistes ou en étroite collaboration avec eux) en fonction des connaissances à extraire et des données disponibles dans différentes sources de données

7. Le mot agile est employé ici au sens de la programmation agile ou *Extreme programming* dont le principe s'écarte résolument du modèle traditionnel de développement logiciel imposant une phase chronophage d'analyse des besoins et de conception avant le développement.

publiques (répertoriées dans BioRegistry) ou privées. Les besoins et l'expertise de l'utilisateur ont donc un rôle central et les tâches répétitives sont automatisées.

Le logiciel MODIM se compose de trois modules : (i) module base de données (i), (ii) module configuration des tâches et (iii) module exécution des tâches. La construction d'un modèle de données par ou avec les biologistes est l'occasion de capturer une partie de la connaissance du domaine sous forme de méta-données et de contraintes d'intégrité. Une fois le modèle relationnel établi, la base de données correspondante peut être créée en utilisant le premier module. Le deuxième module permet de spécifier les tâches qui serviront à peupler la base de données. Chaque tâche possède un type d'entrée unique (par exemple, un identifiant de protéine) et est composée de sous-tâches, chacune dédiée à une source de données (le plus souvent sous forme d'une URL). La configuration d'une sous-tâche consiste à décrire comment reconnaître, dans la source, les données à collecter et où les stocker dans la base de données. La reconnaissance des données à collecter est réalisée par des expressions XPATH (cas de fichiers au format XML) ou des expressions régulières (cas de fichiers texte ou HTML). Une fois terminées, les tâches peuvent être sauvegardées au format XML, éditées et modifiées si nécessaire. L'intégration des données est réalisée par le module d'exécution des tâches. Ce module prend en entrée un fichier ou une requête SQL contenant les données nécessaires à la réalisation de la tâche (par exemple une liste d'identifiants de protéines pour lesquelles on collectera diverses données).

Le logiciel MODIM a été réalisé, sous ma supervision, par Birama NDIAYE grâce à un contrat d'Ingénieur Jeune Diplômé Inria.

Utilisation de MODIM dans trois projets scientifiques : Le logiciel MODIM a été utilisé avec succès dans trois projets distincts :

- Le projet NRPS dont l'objectif est la découverte de nouveaux peptides non ribosomaux et des synthétases associées (ce sont les protéines qui les produisent) [57, 12]. La synthèse non ribosomique est une voie particulière de synthèse de peptides (qui ne passe pas par les phases de transcription et de traduction) rencontrée chez les bactéries et les champignons. Elle est réalisée par d'énormes chaînes de peptides appelées NRPS (pour *Non Ribosomal Peptide Synthetases*) et conduit à la production de molécules actives intéressant divers secteurs industriels (agro-alimentaire, phytosanitaire...). Doris (Database of nOnRIbosomal Synthetases) est une BD de NRPS extraites automatiquement à partir de bases de données généralistes sur les protéines (telles qu'UniProt) ou annotées à l'aide d'outils dédiés. Le logiciel MODIM a permis de collecter et d'intégrer une partie des données sur des candidats NRPS en incluant les génomes de micro-organismes nouvellement séquencés.
- Le projet ICEFinder dont l'un des objectifs est la construction d'une ressource, ICEDb, destinée à faciliter l'identification d'éléments conjuguatifs intégratifs (ICEs) dans des nouveaux génomes de microbes [42]. Les ICEs sont des éléments génétiques modulaires de bactéries qui encodent leur propre excision, transfert, et insertion dans le génome d'autres bactéries. Ces ICEs sont importants à étudier car ils participent à l'évolution des espèces par le biais du transfert horizontal de gènes (pouvant conduire par exemple à une résistance aux antibiotiques). Divers outils d'analyse permettent d'alimenter ICEDb et des tâches MODIM ont permis de collecter des données à partir des séquences de génomes de dizaines d'espèces de bactéries. L'objectif suivant est d'extraire, à partir des ICEs validés par des experts micro-biologistes, des règles pour l'identification automatique de nouveaux ICEs. Cela devrait contribuer à l'annotation automatique des génomes.

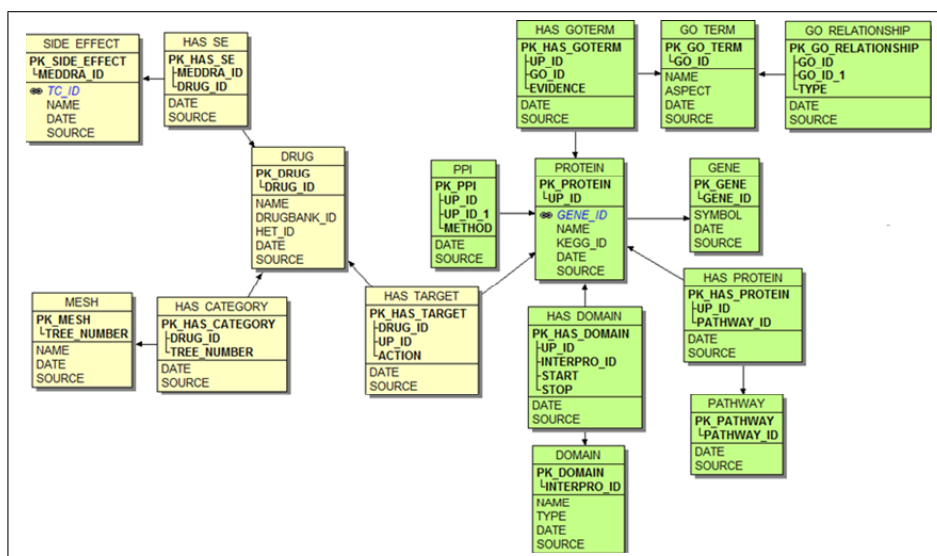


FIGURE 2.2 – Modèle de données de la base NetworkDB (figure tirée de [Bre13]).

- Le projet dédié à l'étude des réseaux d'interaction biomoléculaires pour la compréhension des maladies rares et des effets secondaires des médicaments. Ce projet, correspondant à la thèse d'Emmanuel Bresso, a permis de produire la base NetworkDB [11].

A titre d'exemple, nous présentons brièvement la façon dont MODIM a permis de peupler la dernière ressource citée, NetworkDB, dédiée à l'analyse des effets secondaires des médicaments. Le modèle relationnel correspondant à cette base est présenté Figure 2.2. Les données intégrées concernent les propriétés des médicaments (structure, catégorie ...), leurs cibles qui sont des protéines et les propriétés de ces cibles (fonctions, interactions ...). Quatre tâches MODIM ont été nécessaires pour cette intégration (Figure 2.3). A partir d'une liste d'identifiants UniProt donnée en entrée, la première tâche consiste à interroger la base IntAct et à récupérer toutes leurs interactions protéine-protéine. La deuxième tâche prend en entrée l'ensemble des identifiants UniProt collectés par la première, ce qui inclut les protéines données en entrée et leurs interactants. Cette tâche est divisée en 3 sous-tâches. La première va collecter, de la base Uniprot et pour chaque protéine, le nom de la protéine, l'identifiant du gène et l'identifiant de la protéine dans la base KEGG. La deuxième sous-tâche récupère les termes *Gene Ontology* annotant chaque protéine ainsi que les codes d'évidence associés, et la troisième interroge la base PID afin d'obtenir les identifiants ainsi que les noms des réseaux biologiques. La troisième tâche se concentre sur la collecte des réseaux biologiques KEGG en prenant en entrée les identifiants KEGG des protéines. Enfin, la dernière tâche permet de collecter le symbole du gène disponible dans la base Entrez Gene à partir de son identifiant.

Bilan sur l'intégration de données dans le processus d'ECD : L'approche MODIM est à rapprocher des nombreux efforts fournis pour l'intégration de données biologiques [SSW⁺08]. MODIM est particulièrement utile lorsque les sources de données n'offrent aucune interface programmatoire et doit être vu comme un complément aux autres efforts d'intégration tels que les systèmes BioMart et Taverna. Néanmoins, l'intégration de données en biologie, quelle que soit l'approche adoptée reste un défi à cause de systèmes de nommage pléthoriques qui rendent dif-

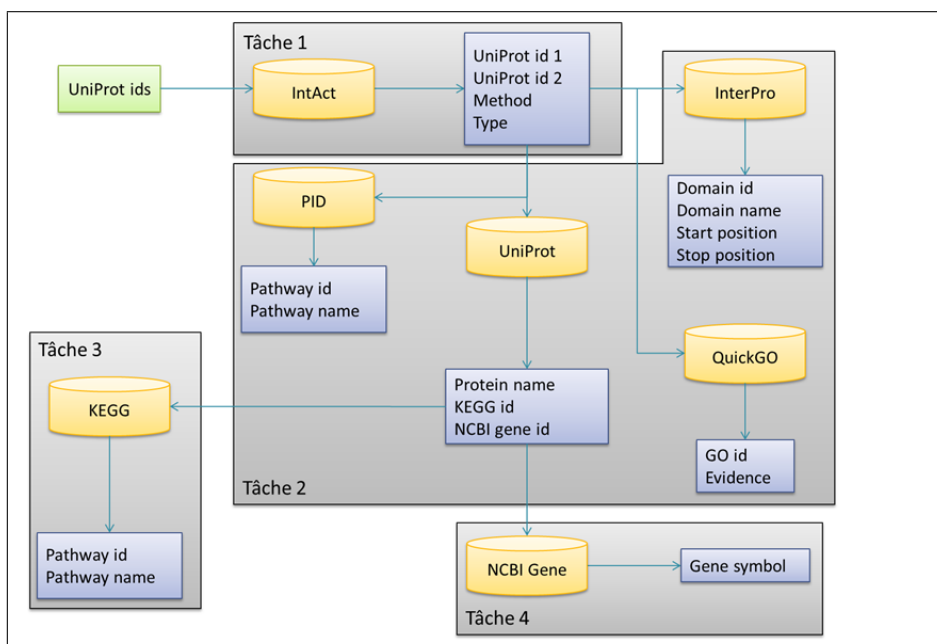


FIGURE 2.3 – Quatre tâches de collecte MODIM pour peupler la base NetworkDB (figure tirée de [Bre13]).

ficile une tâche pourtant basique de l'intégration, la résolution d'identité. De ce point de vue, les initiatives autour des données ouvertes et liées (*LOD*, *Linked Open Data*) sont porteuses de beaucoup d'espoirs [HB11] (je reviendrai sur les *LOD* dans le chapitre 5).

L'approche MODIM avait été initialement motivée par un projet de cartographie du génome humain (avant l'achèvement du séquençage) puis un projet de recherche de gènes candidats pour expliquer une maladie rare pour laquelle le logiciel XCollect avait été développé [28, 26, 23, 65].

En conclusion, quelle que soit la façon d'intégrer les données dans un processus d'ECD, l'expérience que j'ai acquise me conduit à juger important qu'elles le soient dans un SGBD relationnel principalement pour deux raisons. *Primo*, collecter les données en préservant les relations entre divers types d'objets permet d'éviter un appauvrissement des données intégrées (comme cela peut être le cas avec des systèmes de fichiers ou des tableurs) et des difficultés pour manipuler les données et en contrôler la qualité (comme cela peut se présenter en l'absence de toute forme de contrainte d'intégrité). Il est à noter que des données relationnelles ne doivent pas forcément être réduites à une matrice objetsXpropriétés avant d'être fouillées puisque certaines méthodes de fouille sont capables de prendre en compte des données multi-tables. *Secundo*, comme nous le verrons dans le chapitre 4, nous pourrions mémoriser dans le même SGBD les modèles extraits par les programmes de fouille et, ainsi, permettre l'enchaînement avec d'autres analyses et explorations.

2.4.3 Mesure de similarité sémantique entre objets biologiques

La dernière contribution adossée à la préparation des données dans un processus d'ECD est la définition d'une mesure de similarité appelée IntelliGO que nous avons utilisée de deux façons différentes au service de cette préparation. Cette mesure a été développée et testée tout d'abord

par Sidahmed Benabderrahmane au cours de sa thèse que j’ai co-encadrée [Ben11]. Il s’agissait alors de mesurer d’une part la similarité entre deux termes d’une ontologie de domaine structurée en graphe orienté acyclique enraciné ou r-DAG (*rooted directed acyclic graph*) et d’autre part la similarité entre deux objets annotés par des termes (ou concepts) de cette ontologie [7]⁸. L’ontologie qui a donné son nom à la mesure est *Gene Ontology* mais la mesure IntelliGO est adaptée à toute ontologie structurée en r-DAG. La mesure IntelliGO est inspirée de la mesure du cosinus généralisé définie par Ganesan et al., [GGMW03]. Par la suite, Emmanuel Bresso (dont j’ai également co-supervisé le travail) a utilisé cette mesure sur une autre ontologie, le vocabulaire MedDRA, afin de regrouper les termes de sens proche [8].

Préparation à la classification d’objets biologiques

La première utilisation de la mesure IntelliGO a été pour la classification supervisée (clustering) d’objets biologiques annotés par des termes d’une ontologie structurée en r-DAG. En effet, les programmes de clustering (comme présenté dans le chapitre 1) prennent en entrée soit des vecteurs numériques décrivant les objets à classer à partir desquels des (dis)similarités objet-objet sont calculées soit directement la matrice carrée de (dis)similarité objet-objet. En présence de relations sémantiques entre les termes, il n’est pas raisonnable de représenter les objets par des vecteurs de n termes (n est le nombre de termes de l’ontologie) et de calculer les distances euclidiennes ou les valeurs du cosinus en guise de (dis)similarité entre objets. La mesure IntelliGO nous permet donc de calculer une similarité sémantique gène-gène afin de réaliser le clustering de ces gènes. Nous avons ainsi réalisé une analyse fonctionnelle de gènes (sur la base de benchmarks définis pour l’occasion) pour laquelle nous avons obtenu des résultats meilleurs que l’outil DAVID largement utilisé par les biologistes [6, 29].

Une poursuite de cette étude a été motivée par le fait que la mesure de dissimilarité définie sur la base de la mesure IntelliGO (par simple complémentation à 1) n’a pas les propriétés d’une métrique puisqu’elle ne vérifie pas l’inégalité triangulaire. Par conséquent, nous avons exploré la possibilité de réaliser un clustering spectral spécifique (comportant une étape de centrage) [RLKB03]. Nous avons montré une amélioration, qui reste à confirmer, des résultats du clustering [41].

Réduction de dimension d’ensembles de données

Dans le cadre de la thèse d’Emmanuel Bresso, nous souhaitons appliquer des méthodes symboliques de fouille pour analyser les effets secondaires connus des médicaments (molécules chimiques). Un exemple de tâche de fouille est la recherche de motifs fréquents ou de règles d’association parmi les effets secondaires d’une catégorie de molécules (*e.g.*, agents anti-infectieux ou agents cardio-vasculaires) mais aussi la recherche de sous-groupes d’effets secondaires permettant de discriminer deux catégories de molécules. Le dictionnaire bio-médical MedDRA (*Medical Dictionary for Regulatory Activities*), partie de l’UMLS⁹, permet de décrire l’action des médicaments et des vaccins [Res07]. Cette terminologie est en particulier utilisée pour décrire les effets indésirables des médicaments et vaccins dans la base de données SIDER [KCL⁺10].

Lors des expériences de fouille des données décrivant les molécules par les termes MedDRA, nous avons été confrontés au problème de la dimensionnalité des données avec près de 1300

8. Publication jointe à ce document

9. Unified Medical Languages System

termes décrivant quelques centaines de molécules. Présenté comme une structure hiérarchique, MedDRA est en réalité un r-DAG puisque nous avons vérifié que près de 40% des termes ont plus d'un parent direct, que les relations entre termes sont orientées et qu'il n'y a pas aucun cycle. La mesure IntelliGO a été adaptée à MedDRA et le calcul des similarités entre les paires de termes MedDRA a permis de réaliser un clustering des termes MedDRA utilisés pour décrire les effets secondaires en 112 clusters validés par des biologistes. Cela a permis de transformer la représentation des molécules en une forme plus réduite (112 attributs au lieu de 1 300). Nous avons réalisé une étude expérimentale comparative où nous avons montré que cette réduction de dimension conduisait à une terminaison des programmes de fouille et conduisait à des motifs moins redondants et plus faciles à exploiter par un expert biologiste [8].

Dans le même ordre d'idée et dans le cadre de la thèse d'Adrien Coulet (que j'ai co-encadrée), nous avons utilisé des connaissances du domaine pour sélectionner des attributs dans une application de pharmacogénomique [19]. Dans ce travail, une ontologie a été proposée pour formaliser en logique de description les principaux concepts de la pharmacogénomique [17] qui nous a notamment permis d'étudier les relations entre génotypes et phénotypes dans le cadre d'une affection particulière. Cette ontologie permet d'abord de contrôler la collecte et l'intégration de données disponibles dans diverses sources publiques afin d'arriver à instancier une base de connaissances (concepts de l'ontologie augmentés d'assertions). Nous avons également montré comment différents éléments de l'ontologie peuvent être exploités pour réduire la taille d'un ensemble de données à fouiller. Par exemple, au lieu de considérer l'ensemble des variations de certains gènes pour tous les patients, il est pertinent de se limiter aux variations situées dans des régions codantes des gènes ou alors à l'ensemble des variations présentant peu de dépendances fonctionnelles. En termes génétiques, cela correspond au phénomène d'haplotype défini comme un ensemble de variants transmis ensemble lors de la descendance et qui peuvent être résumés par un ensemble réduits de variants appelés tag-SNP. Ainsi, se limiter à l'analyse des tag-SNP permet de découvrir des règles d'association moins nombreuses et surtout moins redondantes [19] (Article joint).

2.5 Conclusion

Découvrir des sources de données pertinentes puis intégrer des données relativement à un problème circonscrit d'analyse est une première étape dans la maîtrise des ressources numériques par les biologistes. Nous avons proposé quelques solutions destinées à aider le biologiste à identifier des sources pertinentes puis à spécifier des tâches automatiques de collecte de données provenant des sources sélectionnées et réalisant leur stockage dans une BD relationnelle.

Afin d'alléger la tâche du biologiste dans le pilotage du processus d'ECD, nous avons également proposé une mesure de similarité sémantique générique et permettant de réaliser une classification non supervisée d'objets biologiques tout en tenant compte de la connaissances encodée dans les ontologies bio-médicales. Cette mesure trouve également une utilisation intéressante pour la réduction de la dimensionnalité d'un jeu de données lorsque cela est pertinent.

Les données biologiques qui sont intégrées le sont ici dans l'objectif de progresser dans la connaissance biologique en passant des données aux connaissances. Il s'agit alors de parvenir à abstraire et à expliciter, à partir d'ensembles de données sélectionnés, des éléments de connaissance (descriptions, règles, contraintes) qui serviront à des raisonnements automatiques et plus généralement à la résolution de problèmes. La logique du premier ordre est fondamentale par

rapport aux bases de données relationnelles et à leurs langages d'interrogation [AHV95]. Cela m'a amenée à m'intéresser aux méthodes inductives capables de gérer des données relationnelles, ce qui fait l'objet des deux chapitres suivants de ce document.

Chapitre 3

Fouille de données biologiques relationnelles

3.1 Introduction

La biologie, la médecine et les sciences de la vie en général, par la complexité et l'importance des problèmes posés, ont suscité de nombreux travaux dans le champ de l'apprentissage automatique et de la fouille de données [WZTS05, BB01, YR12]. Ainsi, pour ne citer que quelques exemples, diverses méthodes de classification ont été adaptées à la recherche de gènes responsables de maladies [MNMHEB12], des algorithmes de découverte de sous-graphes fréquents ont été optimisés pour la fouille de molécules chimiques [DKK05], et des techniques de fouille de texte sont développées pour analyser l'abondante littérature biomédicale au service de diverses applications [KEV05, HPT⁺02, CCA12, ZPZ⁺13].

Une famille de méthodes me semble particulièrement adaptée à la fouille de données biologiques éventuellement en complément d'autres méthodes de fouille : les méthodes de fouille dites relationnelles ou logiques¹⁰ [Rae08, DL01]. Ces méthodes sont présentées dans la section 3.2. Leur application aux données biologiques fait l'objet de la section 3.3. Notre contribution est synthétisée dans la section 3.4.

3.2 La fouille de données relationnelles

Les méthodes d'apprentissage ou de fouille de données présentées jusqu'ici ont en commun de considérer des données propositionnelles, c'est à dire structurées en liste de couples (attribut, valeur). Ces méthodes ne permettent pas de prendre en compte plusieurs types d'entités ou d'objets dotés d'une structure complexe et les relations entre ces objets ou leurs parties. Les méthodes de fouille de données relationnelles permettent d'atteindre ce niveau d'expressivité. Les méthodes de fouille de graphes permettent également d'analyser des données structurées d'un point de vue plus topologique [WM03, AW10].

La fouille de données relationnelles telle que la définit Luc De Raedt (qu'il désigne par *logical and relational learning*), peut également être décrite dans le cadre défini par Mannila et Toivonen pour l'extraction de motifs (introduit dans la section 1.4.1) [Rae08] : c'est la recherche de la

10. Fouille de données relationnelles est la traduction que j'adopte pour *relational data mining*

théorie $Th(L, D, Q)$ définie comme l'ensemble des hypothèses h exprimées dans le langage L qui satisfont la contrainte Q dans l'ensemble d'exemples à généraliser D . L est ici un langage relationnel ou logique et h prend la forme d'une règle ou d'une clause. Un exemple de contrainte est qu'une règle soit couverte par au minimum n exemples. La relation de couverture entre les données et les règles peut être définie à l'aide de l'implication logique. Cette définition sera précisée pour l'apprentissage de concept par programmation logique inductive (Section 3.2.1). Le système ALEPH, que nous avons utilisé, est présenté dans la section 3.2.2. D'autres méthodes pour la fouille de données relationnelles seront passées en revue dans la section 3.2.3.

3.2.1 Apprentissage de Concept par Programmation logique inductive

Préambule (quelques définitions et notations) : Ces définitions sont extraites de [Del88, Llo93]. La syntaxe du calcul des prédicats du premier ordre (CP1) repose sur les notions d'*alphabet* (ensemble de constantes, symboles de fonction, symboles de prédicats), de *termes* (une constante, une variable, ou un terme fonctionnel $f(t_1, t_2, \dots, t_n)$ formé d'un symbole de fonction n -aire f et de n termes t_1, t_2, \dots, t_n), d'*atomes* (formules logiques de la forme $p(t_1, t_2, \dots, t_n)$ ou $\neg p(t_1, t_2, \dots, t_n)$ où p est un prédicat n -aire, c'est à dire une fonction à valeur booléenne, \neg est la négation logique, et t_1, t_2, \dots, t_n sont des termes). Une formule bien formée est composée d'atomes liés par les connecteurs logiques (implication, conjonction notée \wedge , disjonction notée \vee) et peut comporter des quantificateurs (parmi le quantificateur universel et le quantificateur existentiel).

Une théorie clausale est une formule du calcul des prédicats du premier ordre (CP1) que l'on peut écrire comme une conjonction de disjonctions (les quantificateurs universels sont omis) :

$$(l_{1,1} \vee l_{1,2} \vee \dots \vee l_{1,n_1}) \wedge \dots \wedge (l_{k,1} \vee l_{k,2} \vee \dots \vee l_{k,n_k})$$

Chaque disjonction est appelée *clause*. Une clause s'écrit comme une disjonction d'atomes positifs et d'atomes négatifs : $h_1 \vee \dots \vee h_n \vee \neg b_1 \vee \dots \vee \neg b_m$

notée également : $h_1 \vee \dots \vee h_n \leftarrow b_1 \wedge \dots \wedge b_m$

ou encore : $h_1 \vee \dots \vee h_n : \neg b_1 \wedge \dots \wedge b_m$

Une clause de *Horn* (respectivement clause *définie*) est une clause qui a au plus (respectivement exactement) un littéral positif. Un ensemble de clauses définies est un programme logique (langage PROLOG).

Les interprétations sont utilisées pour calculer la valeur de vérité d'une formule. Une interprétation est un modèle pour une formule si cette formule est vraie dans cette interprétation. L'implication logique (traduction de *entailment*) et la satisfiabilité sont définies en utilisant les interprétations. Ainsi F est une conséquence logique de G si tout modèle de F est aussi modèle de G . Une formule F n'est pas satisfiable s'il n'existe aucune interprétation qui soit un modèle de F . Dans le cas où l'on travaille avec des clauses définies, il existe un unique plus petit modèle de Herbrand que l'on choisira pour prouver la satisfiabilité des formules. Ce plus petit modèle est défini théoriquement comme l'intersection de tous les modèles de Herbrand. En programmation logique, il s'agit du plus petit ensemble de connaissances positives (utilisant les termes sans variables construits à partir des constantes et des fonctions qui apparaissent dans les formules) qui permet de satisfaire les formules.

Apprentissage de concept par Programmation Logique Inductive (PLI) : Partons de la définition de l'apprentissage de concept telle que Mitchell l'avait proposée [Mit82] : supposons que sont donnés un langage de concepts L_C , un langage d'exemples L_e , une relation de couverture \in_C qui spécifie comment relier L_C et L_e , et un ensemble d'exemples E d'un concept cible

dans L_C . Chaque exemple est de la forme $(e, Vrai)$ pour un exemple positif ou $(e, Faux)$ pour un exemple négatif. L'objectif de l'apprentissage de concept est alors de trouver une hypothèse H dans L_C qui *couvre* tous les exemples positifs (on dit que H est *complète*) et aucun exemple négatif (H est dite *correcte* ou *consistante*).

La PLI propose un cadre logique pour l'apprentissage de concepts dans lequel [MR94, Rae97] :

- L_C est la logique clausale (CP1 réduit aux clauses),
- H est une théorie clausale (ensemble de clauses de Horn),
- e est une clause de Horn
- l'hypothèse H couvre un exemple e si et seulement si $H \models e$, c'est à dire que e est une conséquence logique de H .

Les systèmes de PLI supposent l'existence d'un ensemble de connaissances a priori B en plus des ensembles E^+ et E^- d'exemples positifs et négatifs du concept à apprendre. L'hypothèse recherchée H doit alors vérifier les conditions de complétude et de correction [DL01] :

Correction :

$$\forall e^- \in E^-, B \wedge H \not\models e^-$$

Complétude :

$$\forall e^+ \in E^+, B \wedge H \models e^+$$

Ce cadre est connu sous le nom d'apprentissage par implication logique (*learning from entailment*). Un autre cadre a été proposé, l'apprentissage à partir d'interprétations. Dans ce cas, H est une théorie clausale, e est une interprétation de Herbrand et H couvre un exemple e à condition que e soit un modèle de H . Ce cadre suppose que les exemples soient spécifiés de façon complète. De Raedt a montré que l'apprentissage à partir d'interprétations est un cas particulier de l'apprentissage par implication logique [Rae97].

L'apprentissage de concept par PLI, comme toute forme d'apprentissage de concept, se ramène à une recherche dans un espace de clauses, des clauses H qui satisfont les propriétés de correction et de complétude. Cet espace est appelé espace de recherche. L'implication logique étant indécidable même dans le cas de la logique clausale, en pratique, la θ -subsumption permet de définir un quasi-ordre sur les clauses définies et de guider le parcours de l'espace de recherche lors de l'induction [Plo70]. C_1 est une généralisation de C_2 par θ -subsumption si et seulement si il existe une substitution θ telle que $C_1\theta \subseteq C_2$. Cette relation de généralisation permet, comme dans la recherche de motifs, d'élaguer des parties de l'espace de recherche. Par exemple si on parcourt l'espace des hypothèses des plus générales vers les plus spécifiques en tentant de satisfaire le critère de complétude, alors il est inutile de tester des hypothèses plus spécifiques d'une hypothèse qui ne couvre pas un exemple positif.

Afin de gérer le bruit inhérent aux données réelles (erreurs dans la description d'exemples, dans la connaissance du domaine, dans l'étiquetage des exemples comme positifs ou négatifs...), il est possible d'alléger les conditions de correction et de complétude d'une théorie en introduisant un critère de qualité et en imposant que l'hypothèse H maximise ce critère [LDB96]. Il est ainsi possible de trouver une théorie clausale couvrant au moins un nombre minimal d'exemples positifs (pouvant donc ne pas couvrir une partie des exemples positifs) et au plus un nombre maximal d'exemples négatifs (pouvant donc couvrir quelques exemples négatifs ou *exceptions*).

Une façon de réduire la taille de l'espace de recherche est la définition de *biais* qui permettent d'orienter le processus d'apprentissage. Grâce aux biais, il est possible de (i) définir un ensemble d'hypothèses intéressantes à considérer sur la base de critères syntaxiques tels que la taille de la clause ou de critères sémantiques tels que les types des arguments dans les littéraux des clauses et leurs modes d'enchaînement ou (ii) de définir un ensemble d'hypothèses interdites.

Il nous faut noter que la portée de la PLI dépasse l'apprentissage d'un concept (ou prédicat) par la découverte de motifs logiques fréquents et discriminants. En effet, une théorie logique est un programme Prolog (ensemble de clauses définies) quelconque et les exemples sont produits par la théorie dans son ensemble et ne correspondent pas seulement à des instances d'un concept cible. L'induction d'une théorie correspond alors à la synthèse de programmes logiques, objectif encore inaccessible. Des solutions ont été proposées pour des versions simplifiées du problème telles que la révision d'une théorie existante ou l'induction de contraintes d'intégrité à partir d'une base de données. Ces travaux sont synthétisés par Luc De Raedt [Rae08].

3.2.2 Le système ALEPH

Plusieurs programmes de PLI ont été développés et certains sont mis à disposition. Le plus utilisé est certainement le programme ALEPH successeur du programme Progol [Sri07]. Ce système est celui que j'ai utilisé dans les projets mentionnés dans ce document. ALEPH est un programme Prolog réalisant l'apprentissage par PLI par implication logique¹¹ introduite par Muggleton [Mug95] et proposant de nombreux paramètres et possibilités de définir des biais et de simuler le fonctionnement de systèmes connus par ailleurs [Sri07]. La théorie est un ensemble de règles qui sont des clauses définies (vide au départ). L'ensemble des graines (*seeds*), tout comme l'ensemble d'apprentissage, est initialisé comme l'ensemble des exemples positifs. L'algorithme de base du programme ALEPH comporte quatre étapes :

1. Sélectionner un exemple à généraliser dans l'ensemble des graines. S'il n'en existe pas, arrêter.
2. Construire la clause la plus spécifique qui implique l'exemple-graine (*bottom clause*) [Mug95]. Cette clause doit vérifier les biais définis (restriction de l'espace de recherche). Cette clause définie comporte généralement de nombreux littéraux.
3. Trouver une clause plus générale que la clause la plus spécifique et qui a le meilleur score par rapport à la fonction d'évaluation qui a été choisie. Cette recherche se fait par couverture séquentielle en partant de la clause la plus générale, la clause *True* (en ajoutant des littéraux de la clause la plus spécifique, en appliquant des substitutions...). Afin de réduire le temps de recherche, la génération des clauses candidates est réalisée grâce à un algorithme *Branch and Bound* [Sri07].
4. Ajouter la meilleure clause trouvée à la théorie. Supprimer les exemples devenus redondants de l'ensemble d'apprentissage et/ou de l'ensemble des graines. Retourner à la première étape.

Divers paramètres servent à configurer différents aspects de la construction d'une théorie dans le programme. Ainsi, la fonction d'évaluation pour le choix de la meilleure règle peut être choisie (par exemple, la différence entre nombre d'exemples positifs couverts et nombre d'exemples négatifs couverts ou la mesure WRAcc présentée dans la section 1.4.1). Il en est de même pour

11. En réalité, ALEPH est, comme le système PROGOL, fondé sur la notion d'implication inverse

le nombre maximal d'exemples négatifs et le nombre minimal d'exemples positifs (et le rapport du second sur le premier) qu'une règle acceptable doit couvrir. Les biais prennent la forme d'un ensemble de déterminations (prédicats pouvant apparaître dans les règles) et d'un ensemble de modes, définissant les types des arguments des prédicats et la manière de les enchaîner dans les règles).

Comme le montre l'algorithme, une itération est réalisée sur les exemples positifs pour trouver la meilleure clause qui généralise la clause la plus spécifique qui implique un exemple (graine) et qui vérifie l'ensemble des biais définis. Lorsque la meilleure règle est trouvée, les exemples positifs couverts par cette règle peuvent être retirés de l'ensemble des graines et/ou de l'ensemble d'apprentissage (qui sert à calculer la couverture d'une règle). Un paramètre spécifique, *induce-type*, permet de définir trois manières distinctes de construire une théorie :

- *induce* : les exemples couverts par une règle sont supprimés des deux ensembles,
- *induce-cover* : les exemples couverts sont supprimés de l'ensemble des graines et pas de l'ensemble d'apprentissage,
- *induce-max* : les exemples couverts ne sont retirés d'aucun des deux ensembles (à l'exception de la graine qui est retirée de l'ensemble des graines).

Les théories obtenues sont sensiblement différentes puisque les théories *induce* et *induce-cover* sont sensibles à l'ordre de présentation des exemples-graines contrairement à la théorie *induce-max* pour laquelle chaque exemple positif est généralisé à son tour. Les règles des théories *induce-max* et *induce-cover* sont plus chevauchantes que celles de la théorie *induce* (par rapport à la taille moyenne de l'intersection des exemples couverts par deux règles).

Les nombreux paramètres et biais font que le programme d'induction par PLI explore seulement une partie de l'espace de recherche en fonction de la valeur de ces paramètres et biais. Les règles logiques qui sont extraites s'apparentent ainsi à des motifs locaux discriminant un sous-groupe d'exemples positifs par rapport aux exemples négatifs. Nous nous appuyerons sur cette constatation pour proposer des prolongements à la PLI (Chapitres 4 et 5).

3.2.3 Extension (*Upgraded*) de méthodes de fouille de données classiques

Les principales méthodes de fouille de données ont été étendues afin de permettre une représentation relationnelle des exemples et des hypothèses à rechercher dans ces données. Ces méthodes étendues sont compilées dans deux ouvrages [DL01, Rae08]. Par exemple, le système WARMR permet l'extraction de règles d'association relationnelles en cherchant les requêtes Prolog (dans un langage contraint par l'utilisateur) qui sont fréquemment satisfaites dans une base de données relationnelles (représentée comme un ensemble de faits Prolog) [DT01]. Luc De Raedt va jusqu'à proposer une méthodologie pour réaliser une extension (dans un des deux cadres logiques d'apprentissage de concept présentés précédemment) d'une méthode de fouille de données propositionnelles afin de prendre en compte un formalisme relationnel en choisissant un langage pour représenter les exemples, un langage pour représenter les hypothèses, en étendant les opérateurs de parcours de l'espace de recherche et le test de couverture d'un exemple par une hypothèse [Rae08].

L'analyse de concept formels, décrite dans la section 1.4.4, a également été étendue pour la prise en compte de relations entre ensembles d'objets, eux-mêmes décrits selon des propriétés binaires dans des contextes formels. Cela a donné lieu à l'analyse relationnelle de concepts (*Rela-*

tional Concept Analysis, ou **RCA**) qui a été définie comme une construction itérative de treillis de concepts jusqu'à la convergence vers un point fixe [HHNV13].

3.2.4 Éléments sur les performances des systèmes de fouille de données relationnelles

L'efficacité et la possibilité de passage à l'échelle est un sujet important pour les programmes de fouille de données et encore plus pour la fouille de données relationnelles [BS03]. Bien évidemment, la complexité du langage de représentation des données et le nombre d'exemples ont un impact direct sur les performances du programme d'apprentissage. Des études théoriques ont été consacrées à l'analyse de la complexité des algorithmes et en particulier à l'évaluation empirique de la θ -subsumption [GSSB00b, GSSB00a]. Ces travaux montrent que la θ -subsumption, consistant à tester si une requête θ -subsume une base de données extensionnelle (l'ensemble des exemples d'apprentissage) est un problème NP-complet. Un cadre d'évaluation emprunté aux problèmes de satisfaction de contraintes (CSP) a permis d'analyser assez finement la complexité de la θ -subsumption en fonction de divers paramètres (nombre de symboles de prédicats, nombre des constantes etc.) et a révélé l'existence d'un phénomène de transition de phase c'est à dire une région étroite dans laquelle le coût de l'opération est maximal. Cette région correspond à des langages de représentation pour lesquels il est difficile de trouver des solutions.

J'ai déjà évoqué des exemples de biais qui permettent de contrôler le processus de génération et de test d'hypothèses (Section 3.2). Des solutions d'approximation synthétisées par DiMaio et Shavlik ont été proposées afin de réduire le nombre de clauses candidates à générer d'une part et le temps nécessaire pour tester les clauses candidates sur les exemples d'autre part [DS04]. A cela s'ajoutent des méthodes heuristiques permettant d'optimiser le couplage entre systèmes de gestion de bases de données relationnelles et programmes d'apprentissage relationnel [YHY06, BDD⁺02].

L'ensemble de ces travaux montrent l'importance du choix du langage de représentation des exemples et l'impact que cela peut avoir sur la possibilité d'une extraction réussie de connaissances. Cela justifie l'importance que nous accordons à l'intégration des données et plus généralement à la préparation des données en lien avec les connaissances du domaine pertinentes vis-à-vis du concept à apprendre. Je reviendrai dans le chapitre 4 sur d'autres possibilités de réduire la complexité de la fouille de données relationnelles.

3.3 Fouille de données relationnelles appliquée aux données biologiques

Comme nous avons essayé d'en rendre compte dans le chapitre 1, les données biologiques comportent une forte composante structurale et relationnelle, conséquence des technologies à haut débit et des progrès de la biologie intégrative. Les méthodes de fouille relationnelles ont donc été naturellement appliquées à ces données et souvent avec succès [Mug99, PC03]. Une des toutes premières applications a porté sur la prédiction de l'activité d'une molécule sur la base de sa structure chimique décrite en termes d'atomes, de liaisons entre les atomes et de groupes fonctionnels (connue sous le vocable *Structure Activity Relationships*) [KSS95]. L'apprentissage de concept par PLI a notamment été appliqué pour :

- identifier des pharmacophores dans le processus de criblage virtuel, c'est à dire des sous-structures tri-dimensionnelles d'une molécule chimique responsables de son activité [FMPS98] ;

- apprendre et prédire différentes classes de structure tridimensionnelle de domaines protéiques [TMS01]. Chaque domaine est décrit par des propriétés globales (*e.g.*, longueur de la séquence, nombre d'hélices...) et des relations d'adjacence entre éléments consécutifs de structure secondaire ainsi que des propriétés de ces éléments (*e.g.*, hydrophobicité moyenne);
- rechercher des molécules diverses du point de vue de la structure et présentant la même activité biologique [TASM08]. Dans ce cas, les distances spatiales entre les atomes composant chaque molécule (active ou inactive) sont prises en compte.

Les programmes de fouille étendus au relationnel ont aussi eu des applications dans le domaine bio-médical. Par exemple, le programme WARMR d'extraction de motifs relationnels a été utilisé pour trouver toutes les sous-structures ayant une fréquence minimale dans un ensemble de molécules [DTK98].

Outre la prise en compte des structures et des relations, la fouille de données relationnelles se prête assez naturellement à la prise en compte de plusieurs sources de données distinctes, chaque source donnant lieu à une ou plusieurs relations (ou prédicats). Ainsi, l'apprentissage de concept par PLI a été utilisé pour la prédiction d'interactions binaires entre domaines protéiques de la Levure (*Saccharomyces cerevisiae*) sur la base des annotations de ces domaines, issues de diverses bases de données publiques (*e.g.*, Pfam, InterPro, PROSITE, UniProt, GO) [NH08]. La PLI a également été utilisée pour l'apprentissage multi-source d'arythmies cardiaques [FQC05].

3.4 Contribution

Nous avons, dans deux études distinctes, formalisé des problèmes d'extraction de connaissances à partir de données biologiques en faisant appel à l'apprentissage de concept par PLI. Ce choix est justifié par l'adéquation du pouvoir d'expression (aussi bien dans le langage des exemples que des hypothèses) aux données et aux connaissances biologiques mais aussi par la possibilité de la prise en compte, lors de l'apprentissage, de connaissances du domaine. Le prix à payer est naturellement (i) la taille de l'espace de recherche et (ii) l'obligation (qui en découle) de définir des biais, ainsi que (iii) la lourdeur de l'application à des données réelles, nécessitant un gros investissement en amont et en aval de la fouille de données. Les progrès combinés dans les processeurs et dans les systèmes opérationnels de PLI permettent de surmonter le premier obstacle. Quant aux deux derniers freins, nous y répondons en définissant un cadre dans lequel le biologiste est assisté lors du déploiement du processus d'ECD.

La première étude que nous avons réalisée concerne la caractérisation de sites d'interaction protéine-protéine à la surface des protéines en partant de la structure 3D de ces protéines (Section 3.4.1) [9]. La deuxième étude concerne la caractérisation et la prédiction de profils d'effets secondaires de médicaments (Section 3.4.2) [11].

Nous représentons l'ensemble des données utiles à l'apprentissage sous forme de clauses définies dans lesquelles les fonctions ne sont pas autorisées. L'absence de fonctions et de récursion constituent une simplification de la logique clausale adoptée dans tous les programmes opérationnels de PLI. La solution adoptée est de représenter une fonction à k paramètres $f(X_1, X_2, \dots, X_k)$ par un prédicat à $k + 1$ arguments $F(X_1, X_2, \dots, X_k, X_{k+1})$ où un tuple $\langle x_1, x_2, \dots, x_k, x_{k+1} \rangle$ du prédicat F spécifie que x_{k+1} est la valeur de $f(x_1, x_2, \dots, x_k)$, cet aplatissement des fonctions n'affectant pas l'expressivité dans la définition des clauses [Rou94]. Dans la première application,

nous avons discrétisé les arguments à valeurs réelles de certains prédicats (par exemple, l'aire de la surface d'une protéine accessible au solvant).

Nous avons également contribué à un environnement logiciel pour le déploiement de la fouille de données relationnelles par extension de l'environnement de découverte de connaissances KNIME¹² (Section 3.4.3).

3.4.1 Caractérisation de sites tridimensionnels d'interactions protéine-protéine

La motivation de cette étude est de progresser dans la compréhension des Interactions Protéine-Protéine (IPP) qui jouent un rôle crucial dans le fonctionnement cellulaire des organismes vivants. Notre travail fait suite aux premiers travaux (mentionnés dans la section précédente) d'extraction de connaissances biologiques explicites pour prédire la structure 3D des protéines. Constatant la disponibilité croissante des structures 3D de protéines, nous avons souhaité réfléchir à l'application de la PLI à la caractérisation de sites tridimensionnels d'interaction protéine-protéine étant donné leur structure. Il existe des méthodes expérimentales pour détecter les IPP mais leur couverture et leur précision sont encore assez faibles et certaines protéines restent difficiles à synthétiser et se prêtent donc difficilement à ce type de méthode de détection.

Afin de pouvoir évaluer la qualité descriptive et prédictive des règles extraites, nous avons choisi de traiter une classe particulière d'IPP : les interactions de phosphorylation, très fréquentes dans la vie de la cellule. Il a été nécessaire de prendre en compte la séquence (structure primaire), la structure secondaire et la structure tri-dimensionnelle des fragments de protéines (désignées également par le terme de *patches*) correspondant à des exemples ou à des contre-exemples de sites d'interaction protéine-protéine lors d'une phosphorylation. La figure 3.1 schématise ces trois niveaux de structure. Les sites de phosphorylation sont répertoriés dans la base de données *Phospho.ELM* mais seuls les sites se trouvant sur des protéines dont une structure 3D existe dans la base *PDB* ont pu être utilisés. C'est sur la base de cette structure que les éléments relatifs aux différents niveaux de structure peuvent être calculés (le fait qu'un résidu se trouve sur une hélice, la surface accessible au solvant d'un patch dans son ensemble ou d'un résidu particulier d'un patch, la distance entre deux résidus...). La figure 3.2 fournit le modèle conceptuel des données collectées et calculées pour cette application. La notation adoptée est celle des diagrammes de classes UML pour représenter les classes (ou types d'entités) et leurs associations.

Nous avons, sur la base de ces données multi-relationnelles, proposé une représentation logique des patches 3D de protéines (tableau 3.1). Un site de phosphorylation a la particularité d'être centré sur un résidu (appelé résidu phosphorylé, en rouge sur la figure 3.1). Nous avons pu exploiter des connaissances du domaine sous la forme d'une catégorisation multiple des acides aminés (ou résidus) qui composent les protéines et de règles d'inférence.

Un exemple de règle obtenue par induction est¹³ :

$$pbs(A) : -p_residue_helix(A, B, +3), high_polarizability(B), p_residue(A, 'Proline', +1).$$

Cette règle exprime qu'un patch tri-dimensionnel A est un site de d'interaction si le résidu à 3 positions (dans la séquence de la protéine) du résidu central du patch se trouve sur une hélice

12. <http://www.knime.org>

13. La syntaxe Prolog est utilisée pour la description des exemples, de la connaissance a priori, et des règles. Selon cette syntaxe, un terme (non *quoté*) dont l'initiale est une lettre majuscule correspond à une variable.

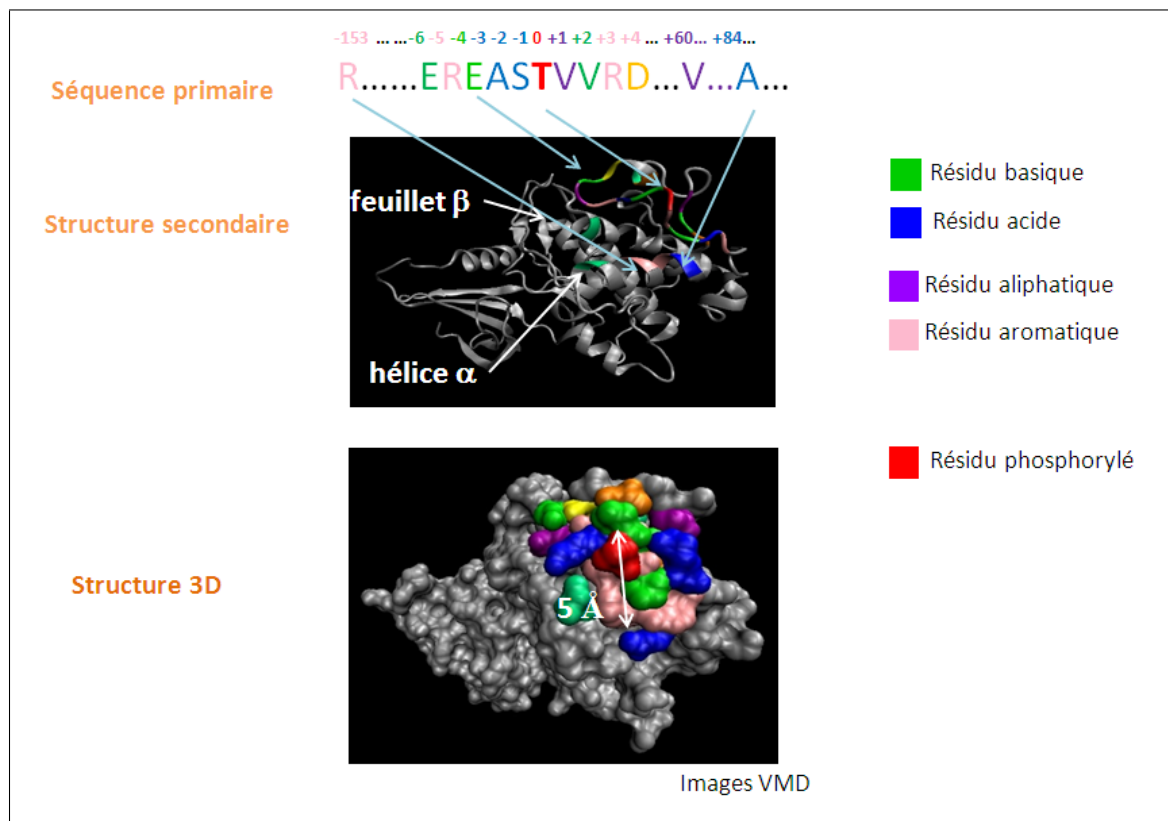


FIGURE 3.1 – Représentation schématique des différentes structures d’une protéine. Le code couleur correspond à une classification particulière des résidus. Les images ont été obtenues avec le logiciel VMD(<http://www.ks.uiuc.edu/Research/vmd>).

Signature(s) de prédicat(s)	Interprétation
$pbs(pid)$	Le patch identifié par pid est un site d’interaction (pbs pour <i>protein binding site</i>).
$p_asa(pid, v)$	v est la surface accessible au solvant du patch pid .
$p_carbon(pid, n)$	n est le nombre d’atomes de carbone (respectivement oxygène, soufre, azote) dans le patch pid
$p_residue(pid, res, pos)$	res est le nom du résidu à la position pos (relativement au résidu central) du patch pid
$p_residue_asa(pid, res, pos, v)$	v est la surface accessible au solvant du résidu situé à la position pos du patch pid .
$p_residue_distance(pid, res, pos, v)$	v est la distance du résidu situé à la position pos au résidu central du patch pid .
$p_residue_helix(pid, res, pos)$, $p_residue_sheet(pid, res, pos)$	Le résidu res situé à la position pos est sur une hélice (resp. un feuillet) dans le patch pid (information issue de la structure secondaire).

TABLE 3.1 – Principaux prédicats logiques pour la description de la structure de patches de protéines.

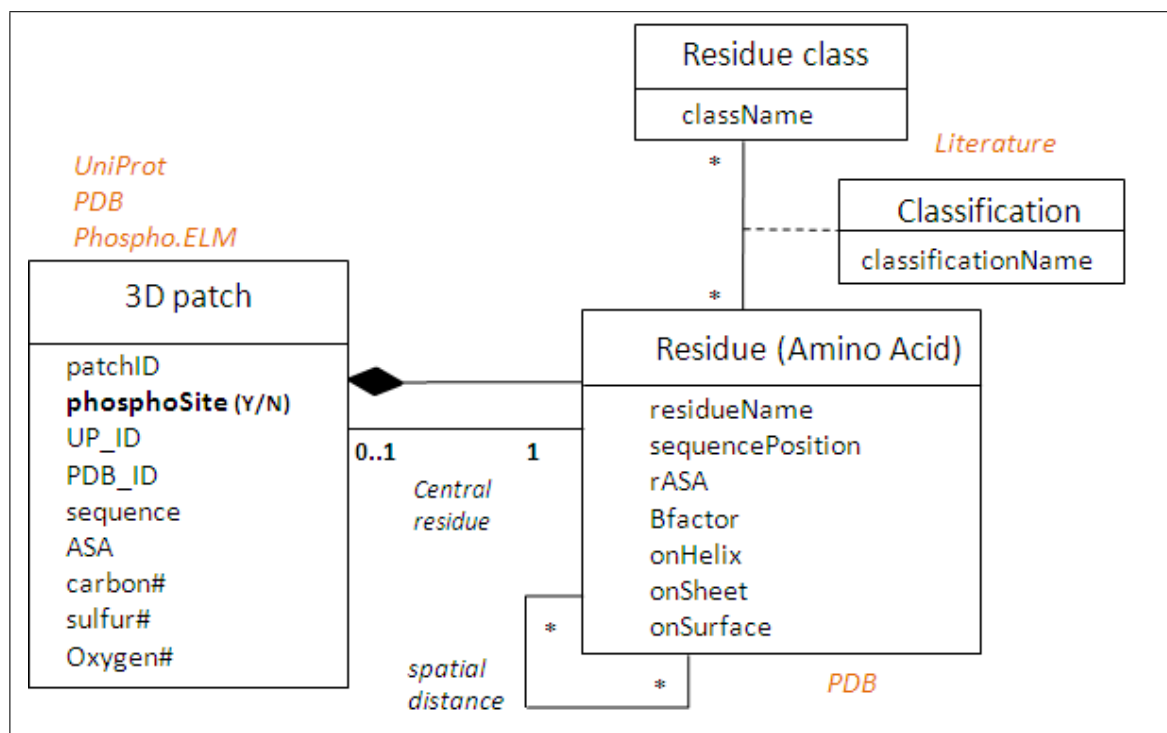


FIGURE 3.2 – Modèle conceptuel des données pour la caractérisation de sites tridimensionnels de phosphorylation. Les noms en italique indiquent les sources de données publiques utilisées.

et appartient à la catégorie des résidus à haute polarisabilité (lys, met, his...) et si le résidu qui succède au résidu central est une proline. Les détails de cette étude sont décrits dans une publication [9].

Cette étude a montré que l'apprentissage par PLI sur des données structurales de protéines permet d'extraire des régularités structurales explicites et intéressantes pour les biologistes et qui complètent avantageusement des systèmes de prédiction sur la base de la séquence primaire tels que kinasePhos¹⁴. Cette étude a également été l'occasion de réfléchir aux moyens d'aider l'expert lors de l'analyse des théories obtenues pour différentes configurations d'un programme de PLI. Je reviendrai dans le chapitre suivant sur l'interprétation, à base d'analyse formelle de concepts, des résultats de la PLI et sur une manière de structurer les résultats de cette analyse dans une base de données inductive.

3.4.2 Définition et caractérisation de profils d'effets secondaires de médicaments

Dans cette étude, réalisée dans le cadre de la thèse d'Emmanuel Bresso, la représentation logique et relationnelle nous permet d'intégrer différents éléments sur les objets que l'on souhaite caractériser et classer, c'est à dire les médicaments (ou molécules chimiques à destination thérapeutique). La motivation de cette étude est de mieux comprendre les effets secondaires des médicaments. Ces effets indésirables représentent en effet la principale raison d'abandon du

14. <http://kinasephos.mbc.nctu.edu.tw/>

processus de développement de nouveaux médicaments au stade des essais cliniques. Les enjeux économiques sont considérables lorsque l'on sait le prix et la durée de ce processus (même si peu d'études s'accordent sur un modèle d'évaluation des coûts, la mise au point d'une nouvelle molécule thérapeutique nécessite en moyenne plus d'une dizaine d'années et plusieurs centaines de millions d'euros). La sécurité dans l'utilisation des médicaments et donc la santé publique est également en jeu lorsque ces effets secondaires ne sont détectés que tardivement (comme cela a été illustré par des scandales récents).

De nombreux projets se sont intéressés à l'étude des effets secondaires des médicaments [LMHX12]. Tous ces projets explorent les effets secondaires isolés et ignorent leurs fréquentes co-occurrences. En effet, tenter de caractériser la classe des médicaments présentant un effet secondaire précis (e.g., nausées) supposerait que cet effet soit isolé et indépendant des autres effets. La réalité biologique fait que, bien qu'un médicament soit conçu pour cibler une protéine en particulier (pour favoriser ou défavoriser certaines actions de cette protéine), ce médicament va inexorablement agir sur d'autres cibles et causer un ensemble d'effets secondaires (céphalée, fatigue...). Autrement dit, il est difficile d'isoler un effet secondaire et de l'étudier.

La figure 3.3 présente le modèle conceptuel des données collectées sur les médicaments et leurs cibles (les attributs des classes ont été omis pour plus de lisibilité). Nous avons, sur la base de ces données, défini la notion de profil d'effets secondaires (PES) pour une molécule comme étant l'association de nombreux effets secondaires présente pour un nombre significatif de molécules. Pour identifier ces PES, nous avons au préalable réduit la redondance et la dimensionnalité du vocabulaire décrivant les effets secondaires des molécules dans la base de données de référence (SIDER). Cela a été réalisé en utilisant la similarité sémantique terme-terme IntelliGO dans le r-DAG MedDRA et a produit 112 clusters de termes MedDRA utilisés pour assigner à chaque molécule une empreinte en effets secondaires. Un PES correspond alors à un motif maximal fréquent (couvrant au moins 10% de l'ensemble des médicaments) extrait de cette matrice d'empreintes.

Chaque PES fait alors l'objet d'un apprentissage de concept par PLI en intégrant des données et des connaissances couvrant l'espace chimique (molécules et ses propriétés) et l'espace biologique (cibles des molécules avec leurs fonctions, leurs processus...). Ces données ont fait l'objet d'une intégration sur la base du modèle de données présenté Figure 2.2. La table 3.2 montre quelques règles de la théorie obtenue pour le profil d'effets secondaires *Dermatitis*. Malgré le caractère bruité des données utilisées, les règles extraites font apparaître l'ensemble des prédicats (définis dans le langage de représentation des exemples) et exploitent correctement les connaissances du domaine. Une évaluation quantitative et comparative des résultats montre que la PLI permet de prédire un PES de façon plus sensible qu'un arbre de décision, lequel présente de bonnes performances dans la détection de molécules ne présentant pas le PES.

Un article décrit cette étude et fait état de cette combinaison de trois méthodes de fouille (clustering sémantique des termes, recherche de motifs maximaux et PLI) [11]. Je reviendrai dans le chapitre suivant sur une manière de structurer les résultats de ces processus successifs de fouille dans une base de données inductive.

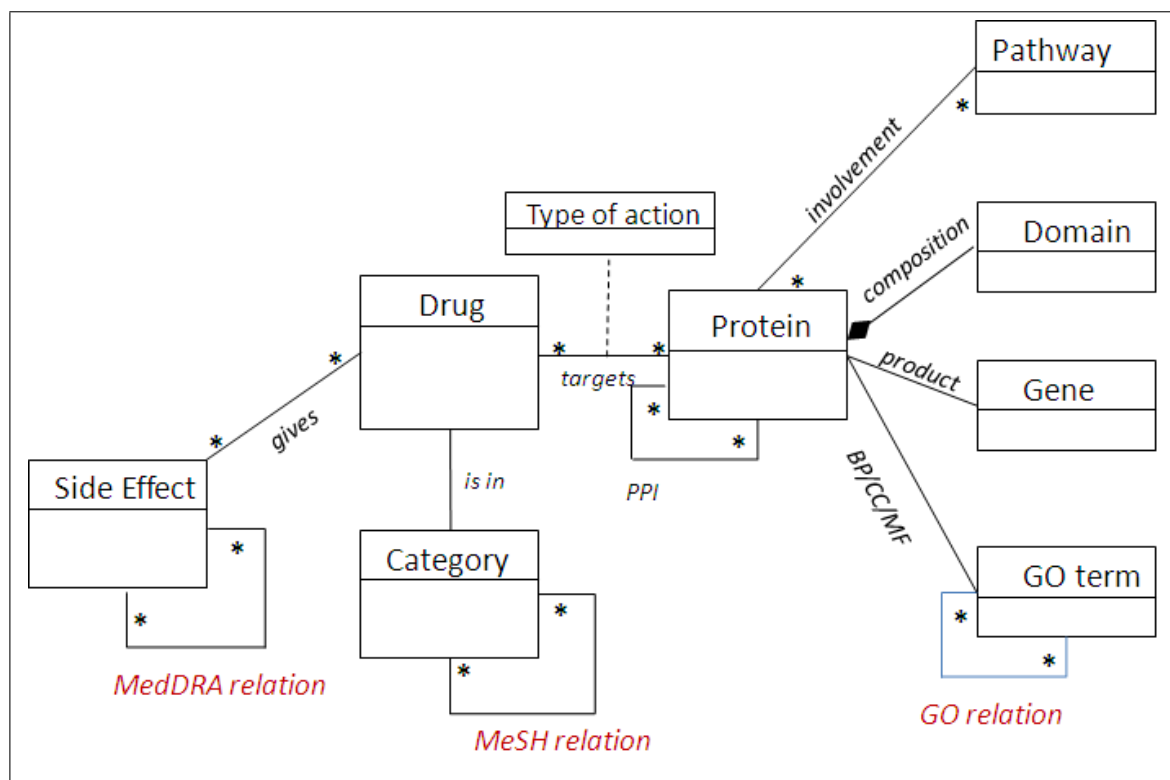


FIGURE 3.3 – Modèle conceptuel des données sur les médicaments et leurs cibles. Les éléments correspondant à des connaissances du domaine sont représentés en italique.

N° règle	Partie condition de la règle	P	N
3	<i>drug_has_target(A,B,'inhibitor')</i> , <i>goterm(B,'cellular response to insulin stimulus')</i>	15	1
18	<i>drug_has_target(A,B,'inhibitor')</i> , <i>goterm(B,C)</i> , <i>go_relation(C,'part_of','Pallium development')</i>	13	1
1	<i>drug_has_target(A,B,'activator')</i> , <i>interact(B,C)</i> , <i>goterm(C,'central nervous system development')</i>	12	1
20	<i>drug_has_target(A,B,'inhibitor')</i> , <i>interact(B,C)</i> , <i>pathway(C,'BCR signaling pathway',pid)</i> , <i>pathway(C,'EPO signaling pathway',pid)</i>	9	1
25	<i>drug_has_target(A,B,'activator')</i> , <i>goterm(B,'lipid binding')</i> , <i>goterm(B,'ligand-dependent nuclear receptor activity')</i>	9	1
6	<i>drug_has_target(A,B,'inhibitor')</i> , <i>goterm(B,'protein homodimerization activity')</i> , <i>drug_cluster(A,'16_gliclazide','hpcc_cluster')</i>	8	0
19	<i>drug_has_target(A,B,'inhibitor')</i> , <i>goterm(B,C)</i> , <i>go_relation(C,'is_a','G-protein coupled amine receptor activity')</i> , <i>drug_cluster(A,'16_Flavoxate','hpcombo_cluster')</i>	8	0
10	<i>drug_has_target(A,B,'inhibitor')</i> , <i>goterm(B,C)</i> , <i>go_relation(C,'is_a','cation channel activity')</i> , <i>goterm(B,'serotonin receptor activity')</i>	7	1

TABLE 3.2 – Règles extraites de la théorie relative au PES *Dermatitis*. P est le nombre de molécules couvertes par la règle et présentant le PES. N est le nombre de molécules couvertes et ne présentant pas le PES.

3.4.3 Environnement logiciel pour le déploiement de la fouille de données relationnelles

Afin de faciliter le déploiement de la fouille de données relationnelles, nous avons, en cohérence avec l'approche MODIM d'intégration de données, spécifié et développé des modules logiciels afin de faciliter la mise en œuvre d'un programme de PLI à partir d'une base de données relationnelles et l'évaluation des résultats de l'apprentissage. Ces modules s'intègrent dans un environnement de découverte de connaissances, KNIME [BCD⁺09]. KNIME est un environnement de découverte de connaissances conçu pour mettre à disposition des implémentations de programmes d'apprentissage ou de fouille avec certaines facilités pour la préparation des données mais aussi pour l'évaluation des résultats de la fouille et même la visualisation de certains résultats. Un environnement de découverte de connaissances permet ainsi de supporter la mise en œuvre du processus d'Extraction de Connaissances à partir de Données (ECD) [3]. Un autre exemple très populaire d'environnement de découverte de connaissances académique est la plateforme WEKA [WF05]. Ces deux environnements permettent de concevoir de façon visuelle et d'exécuter des workflows d'ECD. De nombreux programmes ont été écrits ou intégrés dans ces environnements et correspondent à diverses tâches de fouille : classification supervisée (règles de classification, classification non supervisée, recherche d'associations (motifs fréquents et règles d'association), analyses statistiques (ACP, régressions diverses . . .). Ces deux environnements offrent également un interfaçage avec les systèmes de gestion de Bases de Données (BD) relationnelles les plus répandus afin d'en importer les données à analyser ou d'y stocker des résultats d'analyses.

La seule limitation que nous voyons à ces environnements est l'absence de programme de fouille de données relationnelles. ALEPH est un des programmes de PLI librement accessibles [9] mais qui ne propose qu'une interface en ligne de commande [Sri07]. Nous avons donc dans le cadre du contrat IJD (Ingénieur Jeune Diplômé) Inria de Renaud Grisoni, proposé une extension de la plateforme KNIME avec de nouveaux nœuds afin d'encapsuler le programme ALEPH et de permettre ainsi la fouille de données relationnelles tout en profitant de la librairie de nœuds KNIME. D'autres nœuds ont également été développés pour permettre de réaliser des validations croisées d'un ensemble de règles en logique du premier ordre et d'aider à l'interprétation de ces règles. La figure 3.4 montre un exemple de workflow utilisant certains des nœuds développés. Les programmes (Java) développés pour WEKA ayant été intégrés dans KNIME, cela permet de disposer dans le même environnement des principaux efforts de développement précédents et d'y contribuer.

Ces développements logiciels ont surtout permis de réfléchir aux moyens de faciliter le déploiement d'un processus d'ECD par PLI, qui va de la phase de préparation des données à la phase d'interprétation des résultats de la fouille. Nous avons pu réifier, pour l'apprentissage de concepts par PLI, le cadre conceptuel des BD inductives défini par Imielinski et Mannila [IM96]. Nous proposons de stocker, au même titre que les données collectées, les règles logiques extraites dans le SGBD relationnel hébergeant les données. Nous proposons dans le chapitre suivant des possibilités d'exploitation de ces règles pour prolonger cet apprentissage.

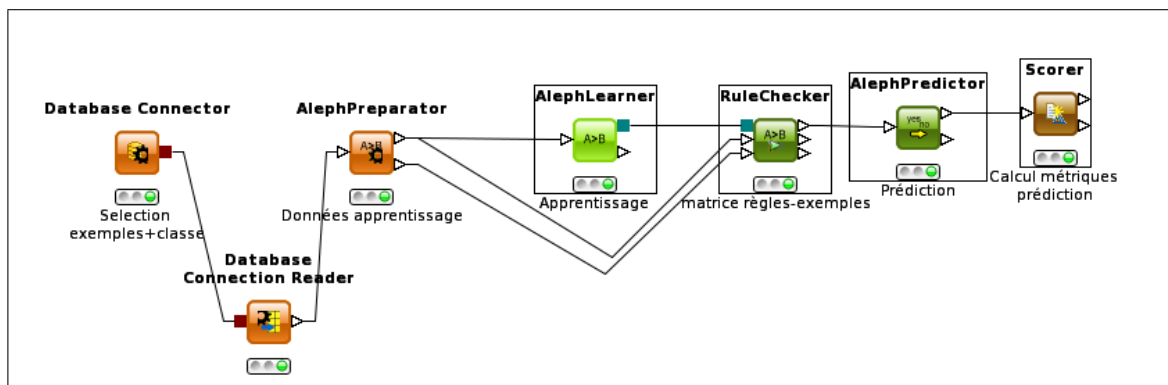


FIGURE 3.4 – Exemple de workflow KNIME.

3.5 Conclusion

Nous avons également eu l'occasion d'appliquer diverses méthodes de fouille à d'autres types des données biologiques notamment dans les domaines de la pharmacogénomique et du criblage virtuel [16] et [18, 38, 36, 37]. J'ai choisi de mettre l'accent sur les méthodes relationnelles pour leur capacité à analyser les données complexes sans sacrifier le pouvoir d'expression, et à intégrer des connaissances du domaine. Comme le souligne Muggleton, la PLI offre la possibilité de produire des connaissances déclaratives par opposition aux connaissances procédurales qu'un arbre de décision ou un réseau de neurones peuvent produire [Mug99]. Une autre application d'apprentissage par PLI que celles mentionnées dans ce chapitre a porté sur la caractérisation explicite et la prédiction des gènes responsables de déficience intellectuelle réalisée dans le cadre d'un deuxième projet d'initiation à la recherche que j'ai co-encadré avec Adrien Coulet. Je reviendrai sur ce projet dans le chapitre 5.

Bien que non fondamental l'effort de développement consenti est important puisque l'objectif est que les biologistes s'approprient l'extraction de connaissances à partir de données (comme cela a été le cas pour les bases de données et les programmes d'analyse) en étant capables de monter et de piloter ces expériences d'ECD.

Enfin, il me paraît important de rentabiliser l'investissement assez chronophage qui est fait dans la phase de préparation des données relationnelles (i) en allégeant le travail des biologistes lors de l'évaluation et de l'analyse des résultats de la fouille et (ii) en permettant d'enchaîner plusieurs itérations du processus d'ECD sur ces données faisant appel à diverses méthodes de fouille. L'intérêt de l'étude des combinaisons pertinentes de méthodes de fouille de données dépasse évidemment le cadre de la biologie. Ces aspects font l'objet du chapitre suivant dans lequel je présenterai quelques combinaisons d'un apprentissage par PLI avec d'autres méthodes.

Chapitre 4

Aide à l'analyse et à l'interprétation des résultats de la fouille : l'ECD dans le contexte des bases de données inductives

4.1 Introduction

Après l'étape de fouille de données, le processus d'ECD se poursuit par une phase d'interprétation et d'évaluation destinée à aider l'analyste à y voir clair dans les résultats de la fouille (Section 4.2). Selon la conceptualisation du processus d'ECD proposée par Fayyad et al., deux issues sont possibles après cette étape (i) soit la réitération du processus d'ECD à partir d'une des étapes précédente (en ré-exécutant le programme de fouille avec d'autres valeurs de paramètres, en exécutant un autre programme de fouille, en testant une autre représentation des données, en réalisant une autre sélection de données...), (ii) soit l'appropriation de ces résultats en vue de leur utilisation pour résoudre des problèmes [FPSS96] (Schéma 1.1). Deux autres cadres conceptuels complémentaires ont été proposés pour décrire le processus d'ECD : celui des bases de données inductives et celui de l'approche LeGo de passage de motifs locaux à un modèle global. Une brève description de ces conceptualisations est donnée dans les sections 4.3 et 4.4. Notre contribution est synthétisée dans la section 4.5.

4.2 Évaluation, validation et visualisation des résultats de la fouille de données

L'évaluation quantitative des résultats d'un programme de classification supervisée consiste à proposer l'estimation empirique la plus précise possible du taux de succès (ou d'erreur) du modèle prédictif produit. La méthode de validation croisée et ses différentes variantes procèdent par prélèvement d'une partie de l'ensemble des données pour le test tandis que le reste sert pour l'apprentissage. Ces méthodes, courantes en statistiques, permettent de produire une bonne estimation de la précision de la prédiction même en cas de données limitées en nombre. D'autres techniques d'évaluation ne sont pas sensibles à la distribution des classes et s'adaptent aux situations où le coût d'une erreur de prédiction n'est pas le même selon que l'on prédit un faux positif ou un faux négatif [PF97]. Par exemple, le calcul de l'aire sous la courbe ROC (AUC)

traçant la proportion des vrais positifs en fonction de celle des faux positifs permet d'évaluer la précision de la prédiction sans prendre en compte la distribution des données. Cette mesure est aussi indiquée lorsque les prédictions peuvent être classées et que ce classement est important.

L'évaluation des résultats d'une classification non supervisée (clustering) est plus délicate. De nombreuses métriques existent pour comparer, sur la base d'une matrice de confusion, deux clusterings réalisés sur le même jeu de données, y compris lorsque l'on a accès à la *vérité terrain* sous forme de benchmark, par exemple [WW07]. Notons que les environnements de découverte de connaissances tels que WEKA et KNIME offrent des bibliothèques qui incluent des fonctions permettant de faciliter l'évaluation quantitative des résultats de la plupart des programmes de fouille de données mono-tables [WF05].

Si les modèles produits sont descriptifs et explicites, il est possible de les évaluer, outre par la précision de la prédiction, selon des critères de nouveauté, d'utilité et d'intelligibilité. Quelques mesures d'évaluation de règles avaient été présentées dans le chapitre 1 (Section 1.4.1). Les biologistes peuvent également jauger la qualité de règles induites en tentant de les rapprocher de connaissances déjà établies. Les règles sont de bonne qualité si elles ne sont pas triviales et qu'elles recoupent ou confirment des résultats publiés dans la littérature ou alors lorsque les biologistes réussissent à établir des liens de causalité dont les règles sont une manifestation.

Lorsqu'ils s'y prêtent, les résultats de la fouille peuvent être visualisés selon différentes modalités afin d'aider l'analyste à mieux les appréhender ou à comparer les modèles obtenus à l'aide de différents programmes, de différents paramètres pour le même programme. . En outre, l'exploration en amont de gros volumes de données d'une part, et certaines tâches de fouille d'autre part, peuvent bénéficier d'une inspection visuelle pour faciliter certaines décisions (sélection de données, choix de programmes, choix de paramètres...). De nombreux travaux ont été réalisés autour de ce qu'il est convenu d'appeler la fouille visuelle de données (*visual data mining*) [Kei02, FGW02]. A un autre niveau, les environnements de découverte de connaissances (tels que KNIME ou WEKA) permettent d'intégrer le processus même d'ECD dans un environnement visuel. Dans ce type d'environnement, un processus d'ECD est représenté comme un workflow, objet visuel et persistant et dont les résultats de chaque étape peuvent être visualisés à tout moment et mémorisés (dans un système de bases de données par exemple).

4.3 Approche LeGo : des motifs locaux aux modèles globaux

L'approche LeGo (*From Local Patterns to Global Models*) est un cadre conceptuel pour modéliser l'extraction de motifs locaux en vue d'une modélisation globale pour diverses tâches de fouille [KCFS08].

LeGo est présentée comme un modèle possible pour l'étape de fouille de données du modèle de processus proposé par Fayyad et al. [FPSS96]. Dans un processus LeGo, l'étape de fouille se décompose en (i) une étape d'extraction d'un ensemble de motifs locaux vérifiant une ou plusieurs contraintes locales (telles que la fréquence, la confiance, le lift) , (ii) une étape de sélection d'un sous-ensemble de motifs locaux vérifiant des contraintes d'ensemble (telles que la diversité ou la non redondance), et (iii) une étape de construction d'un modèle global évalué selon des contraintes globales telles que la précision de la prédiction ou l'interprétabilité. Cette

conceptualisation a permis d'unifier de nombreux travaux antérieurs.

4.4 Bases de données inductives

Dès 1996, Imielinski et Mannila soulignent l'importance de rapprocher le monde des bases de données et de leurs systèmes de gestion avec celui de l'extraction de connaissances et introduisent le concept de *bases de données inductives* [IM96]. L'objectif est de rendre persistants les motifs extraits au même titre que les données afin de permettre leur interrogation voire leur fouille. La solution préconisée pour y parvenir est, par analogie avec les systèmes de gestion de BD relationnelles, (i) la définition d'un langage d'interrogation orienté ECD possédant la propriété de fermeture dont bénéficie l'algèbre relationnelle (qui est à la base du langage SQL) et (ii) la définition d'interfaces de programmation d'applications utiles pour permettre le développement d'applications complexes dans un ou plusieurs langages de programmation cibles.

En réponse à ces attentes, de nombreuses extensions du langage SQL ont été proposées par ajout de constructeurs spécifiques permettant certaines tâches de fouille telles que la recherche de motifs fréquents et de règles d'association. Une étude comparative des principaux langages (MSQL, MINE RULE, SIQL, SPQL, DMX) a été réalisée par Blockeel *et al.*, [BCF⁺10]. Une autre forme d'intégration de la fouille dans les bases de données a été proposée par ces mêmes auteurs comme une extension de la structure de la base de données [FBS06, BCF⁺12]. Cette approche décrit les objets motif ensembliste, règle d'association et arbre de décision comme des vues virtuelles dédiées à la fouille (*virtual mining views*) et fait appel, lors de la matérialisation des vues, à des programmes de fouille de données. L'utilisateur (analyste) se contente alors d'écrire des requêtes SQL sur les tables de la base de données et sur les vues définissant les motifs que l'on peut extraire à partir de chaque table. L'inconvénient de cette approche est la difficulté de représenter sous forme de vues d'autres types de modèles que l'on peut apprendre à partir des données. C'est le cas par exemple des règles de classification en logique du premier ordre.

4.5 Contribution

Notre expérience de l'extraction de connaissances à partir de données biologiques nous amène à tenter d'esquisser une conceptualisation intégrant les principes de l'approche LeGo, du processus d'ECD, et des bases de données inductives.

Nous proposons une conceptualisation des bases de données inductives (BDI) qui n'est pas orientée vers l'intégration de la fouille dans les fonctions du système de bases de données mais vers l'utilisation de ce système comme support pour la persistance des résultats de la fouille, laquelle est réalisée par un programme quelconque. Notre démarche s'apparente à la définition et l'utilisation d'une API permettant de lire les données à fouiller à partir d'une base de données et d'écrire les résultats de programmes de fouille dans la base de données. Pour ce faire, nous avons doté certains noeuds KNIME que nous avons développés de différentes vues qui structurent sous forme de tables les résultats de la fouille ou de l'apprentissage en vue de leur mémorisation dans la base de données. Ainsi, en plus du texte de chaque règle d'une théorie, nous générons la relation de couverture des règles par les exemples d'apprentissage en faisant du *reverse engineering* de chaque règle à l'aide d'un compilateur Prolog. De même le noeud encapsulant le programme de construction d'un treillis de concepts de la plateforme Coron permet de structurer

le contenu du treillis sous forme de tables. Une table définit l'intension de chaque concept, une deuxième table définit l'intension de chaque concept et une troisième table contient les relations de subsumption entre concepts.

Quant à l'approche LeGo, nous en proposons des éléments de réification dans le contexte de la fouille de données relationnelles. LeGo nous permet d'éviter de proposer des solutions *ad hoc* en nous obligeant à abstraire nos processus d'ECD de façon à réutiliser des éléments méthodologiques ou techniques proposés par ailleurs ou à nous positionner par rapport à des travaux connexes.

Ainsi, nous proposons d'étendre le processus d'ECD (figure 4.1) :

- en amont par deux étapes pré-ECD pour constituer la base de données. Il s'agit d'une étape d'identification des sources de données utiles suivie d'une modélisation de ces données et d'une étape de collecte des données à partir de ces sources. Il est à noter que l'annuaire BioRegistry (présenté dans le chapitre 2) est utile pour la première étape tandis que le programme MODIM (présenté dans le même chapitre) l'est pour la seconde ;
- en aval par une étape post-ECD de mémorisation des résultats de chaque étape de fouille dans une BDI. La BDI peut ainsi contenir, en plus des données, des motifs, des règles, des clusters, des treillis ainsi que différentes métriques d'évaluation calculées sur ces modèles.

Ce modèle de processus permet de "faire du LeGo" dans le contexte des BDI en mémorisant les résultats de (i) l'extraction des motifs locaux dans une ou plusieurs itérations du processus et (ii) de la construction du modèle global dans une ultime itération du processus ; la sélection de sous-ensembles de motifs locaux se faisant lors de l'étape de constitution de l'ensemble d'apprentissage de cette dernière itération.

Ce modèle de processus d'ECD itératif et interactif dans le cadre des BDI peut se déployer dans un environnement de découverte de connaissances tel que KNIME en faisant appel à une librairie étendue de programmes pour l'étape d'apprentissage et de fouille. Lors de l'étape de constitution d'un ensemble d'apprentissage, l'analyste peut puiser dans les données pour décrire les objets selon un ensemble d'attributs ou de prédicats mais peut également puiser dans les modèles appris (motifs, règles, clusters) et la relation de couverture (ou d'appartenance à un cluster) pour attacher ces modèles comme propriétés binaires de ces objets (*features*). Cette étape inclut également la structuration des données selon le programme choisi pour la fouille.

Comme nous l'avons déjà précisé, l'évaluation des programmes de classification de données mono-tables fait déjà l'objet de fonctions de la librairie des environnements de découverte de connaissances. Pour les règles apprises par PLI, nous avons étendu un méta-noeud de KNIME afin de réaliser la validation croisée pour deux modèles de prédiction correspondant à une théorie. Le premier, que nous qualifions de naïf, consiste à prédire qu'un exemple est instance de la classe à apprendre s'il est couvert par au moins une règle de la théorie. Le second modèle de prédiction repose sur la construction d'un vecteur couverture pour chaque exemple comme un vecteur binaire indiquant pour chaque règle de la théorie si elle couvre l'exemple. La prédiction pour un nouvel exemple est alors positive s'il existe un exemple d'apprentissage dont le vecteur couverture est distant (respectivement similaire) de moins (resp. plus) d'une distance (resp. similarité) maximale (resp. minimale) fixée. La procédure de validation croisée est par conséquent

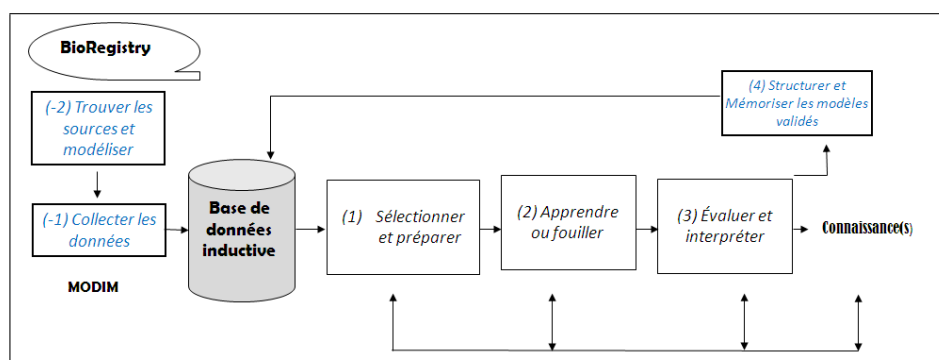


FIGURE 4.1 – Modèle de processus étendu pour l'extraction de connaissances dans le cadre des bases de données inductives.

paramétrée, outre par le nombre de plis (*folds*) et le nombre de répétitions, par la distance ou la similarité et le seuil choisis. Plusieurs mesures ont été implémentées (distance euclidienne, distance de Manhattan, similarité de Jaccard ...).

Nous illustrons dans les deux sections suivantes le processus étendu d'ECD sur la base des deux applications présentées dans le chapitre précédent, à savoir, (i) l'utilisation de l'analyse formelle de concepts comme analyse secondaire à la PLI pour la caractérisation des patchs 3D [10] et (ii) la combinaison de plusieurs méthodes de fouille pour la modélisation des profils d'effets secondaires [11].

4.5.1 L'analyse formelle de concepts comme outil d'interprétation d'une théorie

Dans l'étude de caractérisation de sites d'interaction protéine-protéine à la surface des protéines en partant de la structure 3D, nous effectuons une analyse formelle de concepts (FCA) en aval de la PLI [9]. Cela correspond à deux itérations du processus d'ECD dans le contexte des BDI : la première itération correspond à l'apprentissage par PLI d'une théorie, l'évaluation et la mémorisation de cette théorie (avec la relation de couverture des exemples par les règles) et la seconde itération utilise les règles de la théorie comme de nouveaux descripteurs et réalise la FCA sur la base de contextes formels incluant ces descripteurs.

En considérant les règles logiques obtenues par PLI comme des descripteurs, nous proposons l'utilisation de la FCA avec deux objectifs distincts :

- comme moyen d'aider le biologiste à interpréter les règles d'une théorie en analysant globalement la relation de couverture tout en intégrant d'autres éléments du domaine, susceptibles d'aider le biologiste à raccrocher certaines règles à ce qu'il sait par ailleurs ;
- comme outil heuristique d'exploration de l'espace des configurations d'un programme d'apprentissage définies par une ensemble de paramètres.

Le premier objectif a été exploré lors de notre étude des patchs tridimensionnels de protéines pour caractériser des sites de phosphorylation à la surface des protéines [9]. Nous construisons le treillis de concepts à partir de la relation décrivant la couverture des patchs par les règles (d'une théorie) que nous avons augmentée par l'ajout de propriétés des patchs non utilisées lors de l'apprentissage tels que les domaines fonctionnels sur lesquels se trouvent les patchs. Nous

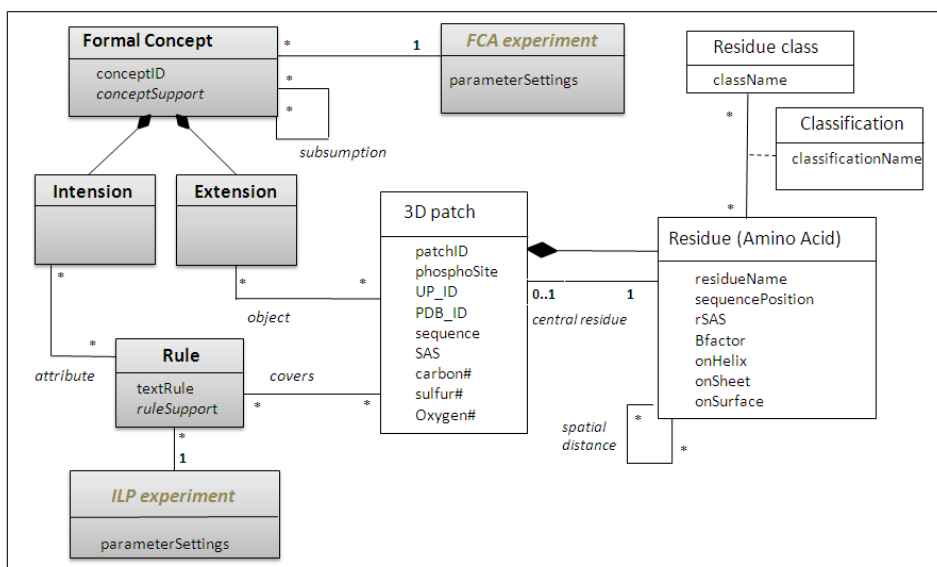


FIGURE 4.2 – Modèle conceptuel des données de la base inductive après deux itérations du processus d'ECD (cf. figure 3.2).

pouvons alors proposer à l'expert d'explorer visuellement le treillis de concepts obtenu ou alors de sélectionner des sous-groupes d'exemples (concepts) couverts par une(des) règle(s) quelconque(s) et ayant une(des) propriété(s) particulière(s). D'autres exemples de requêtes prédéfinies sur le treillis peuvent évidemment être imaginées pour peu qu'elles soient utiles pour aider le biologiste à évaluer la portée des règles trouvées, leur complémentarité...

Le second objectif est utile lorsqu'on souhaite comparer plusieurs théories voire choisir les meilleures règles de plusieurs théories. Nous avons, lors de l'étude des patches 3D, noté que les théories obtenues pour différentes combinaisons de paramètres du programme ALEPH étaient sensiblement différentes ne serait-ce que par la couverture totale de l'ensemble des exemples par les règles d'une théorie. Cette constatation n'est pas surprenante vu les simplifications et les nombreux biais qui poussent l'algorithme de PLI à explorer des parties parfois distinctes de l'espace de recherche. L'étude comparative de théories, à base de FCA sur les données de couverture des exemples d'apprentissage augmentées de connaissances du domaine, requiert des outils heuristiques similaires au premier scénario d'utilisation de la FCA mais aussi des métriques, qui restent à affiner, afin de mesurer la qualité d'un treillis. Par exemple, la mesure de stabilité d'un concept permet d'atténuer l'effet du bruit dans les données sur la formation des concepts et peut se révéler complémentaire au simple support pour la sélection de concepts intéressants à explorer [Kuz07]. La figure 4.2 présente le modèle de données post-ECD pour cette étude où les résultats de la fouille sont structurés en relation avec les données initiales.

L'article décrivant cette étude [9] a été sélectionné pour faire partie d'un ouvrage édité par Springer suite à la conférence jointe IC3K 2012 [10].

4.5.2 Combinaison "à la LeGo" de plusieurs méthodes de fouille au service d'un problème d'apprentissage

À la lumière du schéma de la figure 4.1, la caractérisation d'un profil d'effets secondaires (PES) de médicaments ([11]) peut être décrite comme plusieurs itérations du processus permettant

d'identifier et de mémoriser des motifs locaux en amont d'une itération réalisant l'apprentissage par PLI :

1. Afin de capturer les diverses facettes de la similarité de structure physico-chimique entre molécules, des clusterings ont été réalisés, grâce à des programmes de la société Harmonic Pharma¹⁵), sur la base de différentes représentations des molécules (SMILES, harmoniques sphériques) et de mesures de similarité idoine. Chaque classification donne lieu à un ensemble de clusters et chaque cluster devient un descripteur structurel de la molécule. Autrement dit, chaque clustering donne lieu à la mémorisation d'une relation de couverture des différents clusters par les molécules, exprimant l'appartenance d'une molécule à un cluster.
2. Le choix des meilleurs descripteurs binaires des molécules en termes d'effets secondaires a été effectué grâce au clustering des termes MedDRA réalisé sur la base de la mesure IntelliGO de similarité sémantique terme-terme [7] ;
3. La nouvelle représentation des molécules permet de définir un PES comme un motif maximal fréquent. Cela donne lieu à la mémorisation d'une relation de couverture des différents PESs par les molécules, exprimant le fait qu'une molécule présente un PES.
4. Un ensemble d'apprentissage par PLI est construit sur la base de l'ensemble des molécules présentant chaque PES. Chaque molécule est décrite chimiquement par ses propriétés intrinsèques (telles que sa catégorie mais aussi les clusters structurels auxquels elle appartient) et biologiquement par sa protéine cible et ses propriétés (telles que ses fonctions et ses domaines). Une théorie caractérisant chaque PES est construite séparément.

La précision du modèle de prédiction construit pour la théorie de chaque PES est évaluée par validation croisée (à l'étape (iii) de la dernière itération du processus d'ECD).

Pour finir, la figure 4.3 présente le modèle de données post-ECD après trois itérations du processus d'ECD. Les détails liés aux expériences de fouille (programme utilisé, valeurs des paramètres) ont été omis pour plus de lisibilité.

4.5.3 Conclusion

Utiliser des règles en logique du premier ordre comme des attributs (ou *features*) binaires pour décrire les exemples nous permet de changer de perspective en basculant d'un formalisme de représentation expressif (logique ou relationnel) vers un formalisme plus simple et sur lequel de nombreux algorithmes de fouille peuvent s'appliquer pour prolonger les explorations. Cette idée est similaire à ce qui se fait lors de la *propositionnalisation* dont l'objectif est, pour des raisons de coût de calcul, d'analyser des données relationnelles en réalisant une transformation préalable des données dans un langage propositionnel. Ces travaux, synthétisés par Krogel et al., réutilisent certains mécanismes de l'apprentissage relationnel ou logique pour générer des descripteurs (ou *features*) respectant des contraintes exprimées sous forme de biais avant d'utiliser ces descripteurs comme des propriétés binaires des objets de l'ensemble d'apprentissage [KRZ⁺03]. La limite de la propositionnalisation est le risque de perte d'information dû à la taille exponentielle de l'ensemble d'apprentissage reformulé en langage propositionnel [SR97].

Une autre étude a été réalisée en collaboration avec Yannick Toussaint et des élèves de l'école d'ingénieurs ESIAL (devenue Télécom Nancy) en 2012. Ce travail a porté sur la classification binaire supervisée de recettes de cuisine d'une catégorie d'intérêt. Les données sont issues du

15. <http://www.harmonicpharma.com/>

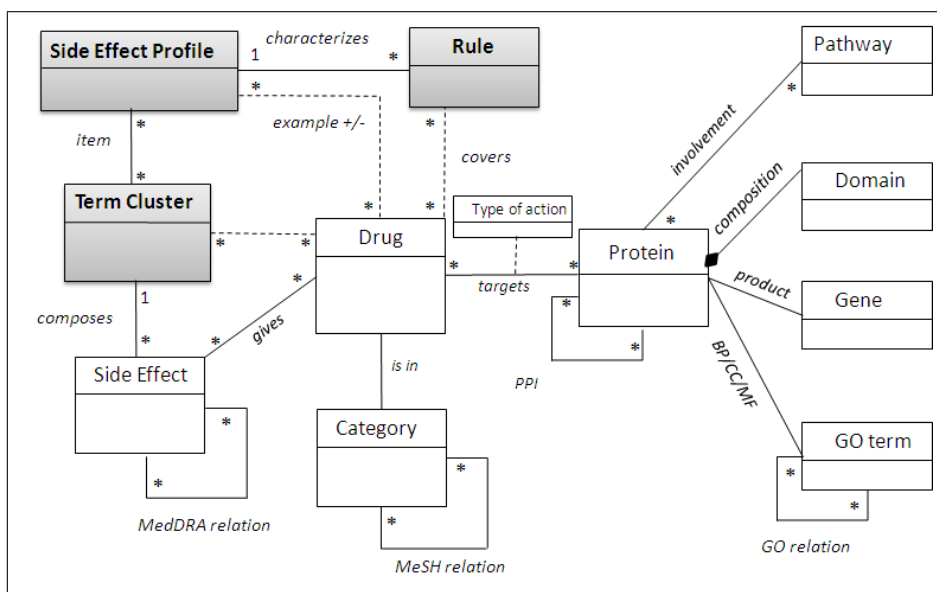


FIGURE 4.3 – Modèle conceptuel des données de la base inductive consacrée aux effets secondaires des médicaments après trois itérations du processus d'ECD (cf. figure 3.3).

projet Taaable, coordonné par Jean Lieber, consacré à l'étude du raisonnement à partir de cas pour l'adaptation de recettes de cuisine aux contraintes de l'utilisateur [BBB⁺08]. Nous avons utilisé le wiki sémantique, wikiTaaable¹⁶, dans lequel les recettes sont catégorisées en plats salés, plats sucrés, plats en sauce, soupes ... [CLM⁺09]. Nous avons appliqué la PLI pour l'analyse des recettes de soupes en exploitant les ingrédients utilisés, les actions réalisées et des connaissances ontologiques sur les ingrédients et sur les actions (issues du wiki sémantique). Un prolongement intéressant de ce travail consiste à exploiter le graphe généré suite à l'analyse du texte d'une recette et qui forme une représentation du déroulement de la recette (une action porte sur un (des) ingrédient(s) dans un état initial et le(s) transforme dans un autre état) [DLBL⁺12]. Ce travail pourrait donner lieu à une combinaison intéressante, selon l'approche LeGo, d'un apprentissage sur les prédicats logiques et d'une analyse des graphes, chaque tâche d'apprentissage donnant lieu à des motifs locaux discriminants d'une catégorie de recette qui serviraient ensuite à proposer un modèle prédictif global d'une catégorie de recette.

Des théories PLI pourraient également être produites selon différentes facettes ou différents points de vue sur les exemples d'apprentissage. Par exemple, nous pouvons rechercher les règles permettant d'expliquer qu'un patch 3D de protéine soit un site d'interaction sur la base de différentes classes de descripteurs dérivés de (i) sa structure 1D (séquence en acides aminés), de (ii) sa structure secondaire, et de (iii) sa structure 3D. Une modélisation globale sur la base des règles obtenues (par construction d'un arbre de décision par exemple) permettrait alors de combiner ces différentes facettes afin de prédire avec précision la classe de chaque patch. Cela s'apparenterait à l'apprentissage dans des univers parallèles dont Berthold *et al.*, proposent une formalisation [BMS07]. Cette approche aurait le mérite dans le cas de l'apprentissage par PLI de faciliter le passage à l'échelle sur des données réelles en travaillant sur des ensembles réduits de prédicats.

16. <http://wikitaaable.loria.fr/>

Énormément de travaux en fouille de données et en découverte de connaissances peuvent être vus comme des combinaisons de techniques et d'algorithmes basiques connus [FPSS96]. Nos travaux ne forment pas une exception à cette tendance car proposer des nouveaux algorithmes est important mais il me semble tout aussi important de faciliter le déploiement d'applications d'ECD en favorisant la réutilisation des efforts de programmation et d'optimisation antérieurs. Nous avons ainsi proposé deux modes de coopération entre un apprentissage par PLI et d'autres méthodes de fouille dans le cas de deux applications d'ECD distinctes (analyse formelle de concepts après la PLI et coopération de divers clusterings et de recherche de motifs ensemblistes pour préparer la PLI). Au delà des données et des problèmes biologiques, j'ai proposé une abstraction de nos travaux afin de proposer un modèle possible de processus d'ECD dans le cadre des bases de données inductives compatible avec les principes de construction d'un modèle global à partir de motifs locaux extraits au préalable.

Chapitre 5

Bilan et perspectives de recherche

5.1 Bilan

Les données et les problèmes biologiques par leur complexité restent un défi pour l'extraction de connaissances. Une fois le problème d'apprentissage bien posé, se pose la question du langage de représentation à utiliser, de la forme des connaissances à extraire et du processus d'Extraction de Connaissances à partir de Données (ECD) adéquat. En amont de la fouille de données, il est admis que la préparation des données représente la majeure partie (en temps) d'un processus d'ECD lorsque le programme de fouille est déjà prêt. En aval de la fouille de données, il n'est pas raisonnable de proposer aux biologistes des résultats de fouille parfois aussi volumineux, voire plus, que les données initiales. Notre contribution dans ce contexte a été de proposer quelques solutions pour ces étapes clés du processus d'ECD tout en exploitant les connaissances du domaine lorsqu'elles sont disponibles.

Néanmoins, la panoplie des problèmes d'extraction de connaissances à partir de données biologiques est large et les approches que nous avons adoptées ne sont pas suffisantes pour prendre en compte toute sa diversité. C'est le cas, par exemple, des phénomènes continus qui sont importants lorsqu'on cherche à modéliser des systèmes biologiques (le niveau de transcription d'un gène, la quantité nécessaire d'un réactif...). De nombreux travaux actuels ont pour objectif cette modélisation selon différentes approches. Des études étendent l'apprentissage relationnel à l'apprentissage relationnel statistique (*Statistical Relational Learning*) ou PLI probabiliste pour prendre en compte l'incertitude dans les données [RK03, GFKT01]. Certaines de ces méthodes s'appuient sur des noyaux qui permettent de définir des similarités sur des données structurées, des graphes, et des données relationnelles [Gö3, Rae08]. La puissance des méthodes à noyaux repose sur la possibilité, grâce à la définition d'une fonction noyau valide et adéquate, de réaliser un plongement de données complexes dans un espace linéaire doté d'un produit scalaire rendant possible l'application des méthodes d'analyse classiques telles que le clustering. Des applications ont concerné notamment l'inférence de réseaux de régulation de gènes [GTDdB07]. Comme cela apparaîtra dans mes perspectives de recherche, il paraît judicieux de combiner ces méthodes à noyaux avec d'autres méthodes (notamment symboliques) afin de concevoir des processus d'ECD capables de gérer les différents aspects d'un problème biologique (Section 5.2.2).

5.2 Projet de recherche

Mes perspectives de recherche sont nombreuses et j'en présente quatre. La première perspective, à court terme, porte sur la construction de modèles prédictifs précis en aval de la PLI (Section 5.2.1). La deuxième perspective, à moyen terme, concerne l'introduction d'un nouveau type d'apprentissage, l'apprentissage transductif, qui me semble complémentaire de l'apprentissage par induction pour extraire des connaissances biologiques de meilleure qualité (Section 5.2.2). La troisième perspective, à moyen terme également, porte sur l'intégration des données ouvertes et liées et l'étude de l'impact sur le processus d'ECD (Section 5.2.3). A plus long terme, la dernière perspective concerne la conceptualisation et la mise en œuvre d'un environnement d'analyse de données patients dans leur diversité avec l'ambition de contribuer à la médecine personnalisée (Section 5.2.4).

5.2.1 Prédicteurs en aval de la programmation logique inductive

Nous prolongeons les travaux antérieurs en étudiant trois types de prédicteurs que l'on peut construire à partir d'une théorie apprise par PLI :

1. un prédicteur naïf,
2. un prédicteur à base de distance (ou de similarité) de vecteurs couverture
3. un prédicteur global

Les deux premiers types de prédicteurs ont été présentés dans le chapitre 4. Une troisième façon de construire un modèle prédictif sur la base d'une théorie est de construire un modèle de classification global (arbre de décision, SVM...) sur la base de la relation binaire de couverture des règles de la théorie par les exemples de l'ensemble d'apprentissage à laquelle on ajoute l'information sur la classe (colonne supplémentaire). Il est alors possible d'estimer la qualité du modèle de prédiction global obtenu. Il est intéressant à ce niveau, d'explorer les moyens de sélectionner parmi les règles issues d'une ou plusieurs théories, les règles locales qui vont donner le meilleur modèle prédictif global. Une idée intéressante serait d'attribuer un score à chaque règle en fonction de sa contribution au modèle global afin d'utiliser les règles extraites par PLI comme de nouveaux descripteurs pour construire un modèle global. Davis *et al.*, ont mis en œuvre cette idée dans l'approche SAYU (Score As You Use) dans laquelle les règles sont précisément extraites par PLI et le modèle global est un réseau bayésien. Notre objectif est d'effectuer la sélection des meilleures règles par la construction incrémentale d'un prédicteur global et d'évaluer le gain en termes de précision de la prédiction.

Dans l'étude des profils d'effets secondaires (PES), un problème intéressant se pose lors de la fusion des théories obtenues pour les différents profils. Comment prédire un ou plusieurs profils pour une nouvelle molécule sur la base des matrices de couverture des règles de chaque théorie par les exemples ? On pourrait simplement construire un arbre de décision multi-classe, une classe correspondant à un PES ou à une combinaison de plusieurs PES puisqu'une molécule peut élargir à plusieurs profils. Une alternative serait de construire une théorie multi-classe (une classe par PES) au lieu d'une théorie par classe, des travaux ont d'ailleurs été consacrés à l'adaptation de la PLI pour la construction et l'évaluation de théories multi-classes [AF10b, AF10a]. Néanmoins, les classes correspondant à PES_i et PES_j ne sont pas indépendantes puisqu'une molécule peut élargir à plusieurs PES. Par conséquent, les solutions d'un arbre de décision ou d'une théorie multi-classe ne sont pas pertinentes puisque nous sommes ici en présence d'un

problème de classification multi-label pour lequel des méthodes existent et peuvent être adaptées à notre problème [DWCH12, ACMG13]. Une autre piste à explorer est de réaliser la prédiction des profils d'une molécule grâce à plusieurs tâches d'inférence transductive à partir d'exemples étiquetés et décrits par leur vecteur couverture et sur la base d'une mesure de similarité adéquate entre deux vecteurs couverture, une tâche par PES (je reviendrai sur la transduction en Section 5.2.2). Cette solution devra être comparée notamment avec la solution qui s'appuierait sur autant d'arbres de décision (ou tout autre modèle prédictif global) que de PES.

Nous explorons des questions similaires dans un projet concernant la caractérisation des gènes responsables d'un phénotype. Cela correspond à une étude démarrée à l'occasion d'un projet d'initiation à la recherche proposé à des élèves de Télécom Nancy en collaboration avec Adrien Coulet [56]. Afin d'améliorer la qualité descriptive aussi bien que prédictive des règles et dans l'esprit de l'apprentissage dans des univers parallèles [BMS07], nous définissons différents sous-espaces de descripteurs des gènes et de leurs produits (annotations fonctionnelles, données structurales et d'interaction...). Les possibilités de production de règles et de combinaison de ces règles sont nombreuses. Nous testons actuellement la qualité du modèle global de prédiction obtenu selon différents classifieurs (arbre de décision, SVM, réseau de neurones) par simple fusion des règles obtenues par PLI dans deux sous-espaces différents de descripteurs. Nous souhaitons appliquer les méthodes d'ensemble telles que le *bagging* et le *boosting* adaptées à la PLI afin de construire des modèles de prédiction encore plus performants [Qui96, dCDPCS02].

Du point de vue de l'intégration des données, cette étude relative à l'étude des gènes responsables d'un phénotype nous permet d'amorcer en douceur le virage vers les données ouvertes et liées (DOL ou *LOD*, *Linked Open Data*) [BHBL09, HB11, 56]. Elle se poursuit dans le cadre d'un projet PEPS Mirabelle (appel conjoint CNRS et Université de Lorraine) porté par Adrien et destiné à explorer l'utilisation des DOL pour la découverte de connaissances (je reviendrai sur ce sujet dans la section 5.2.3).

5.2.2 Introduction de la transduction dans le processus d'ECD

Jusque là, nous avons vu le processus d'apprentissage comme (i) un processus d'induction qui permet de généraliser un ensemble d'instances étiquetées (par la classe) pour produire un modèle en prenant en compte, si possible, des connaissances *a priori* suivi (ii) d'un processus de déduction qui permet d'inférer (prédire) une étiquette sur la base du modèle appris pour toute instance de test. Néanmoins, sur la base des différentes applications d'ECD que nous avons réalisées à partir de données biologiques, nous constatons que, malgré le caractère souvent pléthorique de ces données, lorsque l'on s'intéresse à un problème d'apprentissage précis, il n'est pas aisé de disposer d'autant d'exemples (ou instances étiquetées) d'apprentissage que l'on souhaiterait. Plus précisément, il peut être coûteux d'étiqueter des instances alors même que l'obtention des instances de test ne l'est pas. C'est le cas par exemple lorsqu'on cherche à prédire l'activité d'une molécule sur une cible : une molécule est dite active si une expérience a montré qu'elle se lie à la cible alors que les molécules à tester proviennent de chimiothèques volumineuses [WPCB⁺03]. Dans ce type de situations, la faible taille de l'échantillon d'apprentissage et le bruit inévitable dans les données biologiques font qu'il est assez difficile de construire un modèle prédictif global précis. C'est précisément dans ces cas que l'inférence transductive s'applique pour prédire directement les étiquettes des instances de test à partir des instances étiquetées [GAV98, CSZ06, CM07]. Vapnik avait justifié l'intérêt de ce type d'inférence par le fait que l'apprentissage supervisé par induction (ou généralisation) oblige, pour prédire la classe pour un ensemble de test, à résoudre

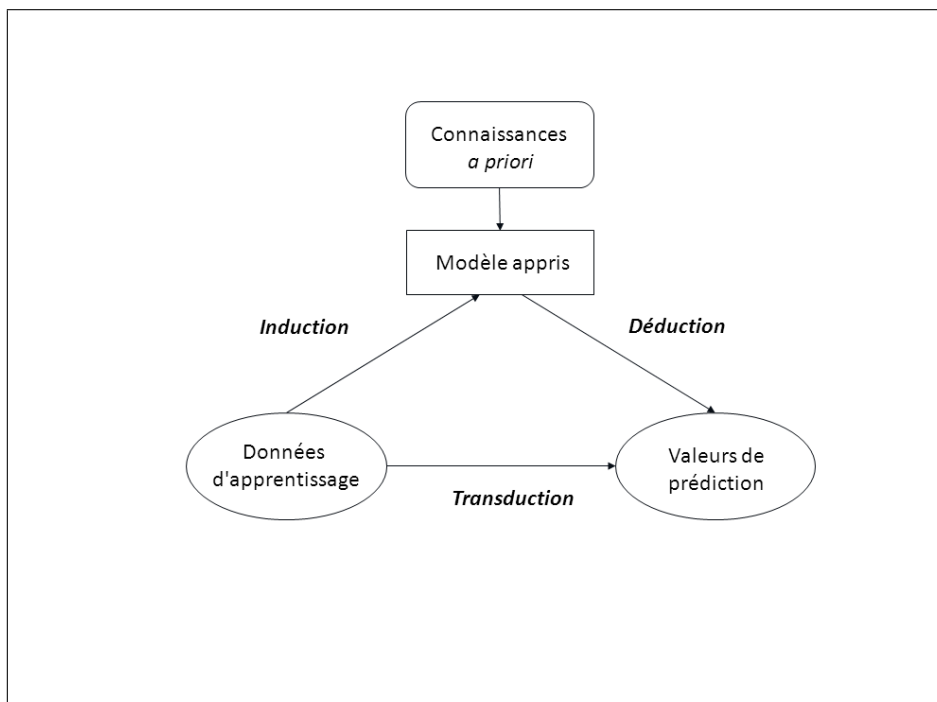


FIGURE 5.1 – Deux types d'inférences : induction-dédution et transduction (figure adaptée de[CM07])

un problème plus difficile qui est l'induction d'une fonction générale qui prédit la classe pour n'importe quel nouvel exemple sur la base de l'ensemble des exemples étiquetés [Vap99]. La figure 5.1 schématise les deux types d'inférences tels que présentés par Cherkassky et Mulier [CM07].

L'hypothèse qui est faite lors de l'inférence transductive est que la fonction d'étiquetage (correspondant au modèle) est continue au moins localement et il s'agit de propager les étiquettes des instances qui en sont dotées aux instances de test. Ainsi, la transduction impose deux contraintes : les instances de test doivent être connues et la fonction de prédiction n'est pas produite. Un algorithme très simple de transduction est inspiré des méthodes de clustering par partitionnement et consiste à partir d'un ensemble d'instances étiquetées (par k classes) et d'instances de test à (i) partitionner (par une méthode de clustering) toutes les instances en n sous-ensembles tels que toutes les étiquettes des instances de chaque sous-ensemble soient identiques et (ii) propager dans chaque sous-ensemble l'étiquette aux instances de test. Des travaux ont développé et appliqué le principe de l'inférence transductive avec des méthodes à noyaux, notamment sur la base de représentations sous forme de graphes et sur des problèmes tels que la classification de séquences de protéines [CSZ06].

Je souhaite explorer les possibilités d'inférence transductive sur des données bio-médicales complexes. J'envisage deux scénarios de complémentarité entre méthodes symboliques et numériques d'analyse de données. Dans le premier scénario, la transduction permettrait d'agrandir la taille de l'ensemble d'apprentissage et donc d'améliorer la précision de l'apprentissage symbolique qui s'en suivrait. Dans le second scénario, l'apprentissage symbolique (extraction de motifs ensemblistes, motifs séquentiels, de règles d'association, de règles de classification) permettrait de générer des motifs locaux qui peuvent servir à décrire (totalement ou partiellement) les exemples

et à faciliter la transduction. Un problème crucial dans les deux scénarios envisagés est la définition d'une mesure de similarité qui capture bien la sémantique des données et qui sera à la base de l'inférence transductive. Pour cela, nous pourrions adapter des mesures de similarité sémantiques existantes [7, PFF⁺09]. Plus généralement, il est utile d'exploiter les résultats des travaux proposant de définir des noyaux prenant en compte des données structurées ou relationnelles pour permettre le calcul de similarités [Gö3].

Une application (micro)biologique pour laquelle l'inférence transductive serait pertinente est l'aide à l'annotation de génomes bactériens et plus précisément l'identification d'éléments conjuguatifs intégratifs (ICE) [42]. Un ICE est une partie d'un génome bactérien doté d'une structure multi-modulaire complexe et qui joue un rôle important dans le transfert horizontal d'éléments génétiques entre bactéries (et qui a pour conséquence l'évolution des bactéries, par exemple pour devenir résistantes aux antibiotiques). Un travail préliminaire très chronophage et fastidieux de bio-analyse, d'annotation experte et de collecte de données permet actuellement d'étiqueter quelques dizaines d'ICEs dans une cinquantaine de génomes récemment séquencés. La transduction permettra de propager les étiquettes, sur la base de mesures de similarité qui restent à définir, aux séquences de centaines de nouveaux génomes bactériens. La découverte d'enzymes de synthèse de peptides non ribosomiques, aussi appelées NRPS (de l'anglais *non-ribosomal peptides synthetases*) est un autre exemple d'application assez analogue à la recherche d'ICEs [12].

Une troisième application serait la pharmacovigilance afin de prédire l'arrivée d'un événement non désirable pour un patient lors d'un suivi médical (Section 5.2.4).

5.2.3 Intégration des données biologiques ouvertes et liées dans le processus d'ECD

Les efforts dans la publication de données structurées et faciles à relier les unes aux autres ont contribué à l'essor de ce qu'on appelle le web des données. En particulier, l'expression *Données Ouvertes et Liées* (*Linked Open Data*, DOL) fait référence à un ensemble de bonnes pratiques visant à publier et relier les données structurées sur le web en utilisant les technologies du web sémantique (langages RDF, OWL, SPARQL) [BHBL09]. De nombreux fournisseurs de données adoptent petit à petit ces bonnes pratiques et donnent ainsi accès à un espace global de données comportant déjà des milliards de triplets RDF [HB11]. C'est en particulier le cas des données biologiques à travers les projets Bio2RDF [BNT⁺08] et RDF Platform de l'institut européen de la bio-informatique (<http://www.ebi.ac.uk/rdf/>). Néanmoins, cette évolution pose au moins deux défis à relever avant de pouvoir exploiter ces données : (i) les moyens d'interrogation globale des données liées et (ii) leur intégration dans une source unique. Notre projet porte sur l'intégration dans le processus d'ECD des données biologiques ouvertes et liées. L'intérêt naissant pour cette problématique est illustré par deux workshops dédiés à la fouille de données ouvertes et liées, en marge des conférences ECML/PKDD 2013 à Prague et WWW 2014 à Séoul.

Il n'est pas aisé pour les biologistes qui souhaitent intégrer des données pour une expérience d'ECD de désigner précisément le sous-ensemble des données pertinentes. En effet, les DOL, quoique ouvertes et flexibles, posent le problème, lorsqu'elles sont explorées comme un gigantesque graphe, de la désorientation de l'utilisateur qui cherche une information (comme cela avait été le cas lors de l'apparition des documents hypertextes et avant l'arrivée des moteurs de recherche). Le projet Sindice.com, en cours de développement, a pour ambition d'offrir une porte d'entrée unique au web sémantique [ODC⁺08]. De façon analogue aux moteurs de recherche pour

la recherche dans le web traditionnel, il est possible de soumettre une requête simple (mots-clés ou URI) ou une requête SPARQL. Néanmoins, cet index étant généraliste, nous aurons les mêmes limitations que nous avons rencontrées avec les moteurs de recherche pour trouver efficacement les sources de données biologiques. Comme nous l'avons déjà mis en œuvre avec l'annuaire Bio-Registry et l'approche MODIM, nous privilégions une cartographie des sources de DOL et une modélisation des données à collecter précédant la collecte elle-même. Cette modélisation permettrait de capturer le plus précisément possible la sémantique du domaine ciblé (dans lequel on cherche à extraire des connaissances). Vu la structure d'un triplet, associant une propriété à un objet, le modèle conceptuel entité-association (ou de façon équivalente un diagramme de classe UML) semble être le meilleur moyen de capturer la sémantique des données à collecter. Nous pourrions nous appuyer sur des travaux antérieurs de formalisation des correspondances entre un modèle de données et les logiques de description comme langage de représentation des connaissances, au dessus du langage RDF [BBJN96, HPT01, PLC⁺08]. L'approche proposée par Poggi et al., consiste à concevoir une ontologie dans une logique de description particulière (partie intensionnelle) puis à définir des *mappings* qui servent à instancier l'ontologie (partie extensionnelle ou assertionnelle) à partir de données provenant de sources hétérogènes afin d'intégrer ces sources de données et en permettre l'interrogation à travers l'ontologie [PLC⁺08]. Pour notre part, à l'inverse, le modèle de données représentant le domaine d'intérêt nous servirait de vue sur les DOL et des wrappers pourraient être définis sous forme de requêtes SPARQL adressées à différentes sources de DOL. Dans les DOL, un serveur de données (SPARQL Endpoint) fournit des données relatives à des objets identifiés par des URI reliés par des relations elles-mêmes identifiées par des URIs. Cartographier les DOLs biologiques reviendrait à typer les objets et les relations, grâce à des mécanismes génériques, relativement à une ontologie universelle (à créer si nécessaire) et à structurer sous forme d'annuaire de méta-données le contenu des serveurs de DOL. Sur un plan opérationnel, nous devons nous appuyer sur les efforts de développement faits dans le cadre du projet Bio2RDF qui exploitent une ontologie biologique intégrée (SIO, pour *Semantic science Integrated Ontology*) pour indexer et résumer le contenu des sources de données mais aussi documenter la provenance et la qualité de ces sources [CCTAD13].

Quelles que soient les solutions d'intégration qui seront retenues, il me semble important de ne pas rompre le lien avec les (systèmes de gestion de) bases de données relationnelles. En effet, malgré les promesses des DOL et le fait que leur intégration soit plus aisée (par construction), il n'en demeure pas moins que la couverture des DOL est encore relativement faible et que ni la qualité ni la consistance ne sont garanties [BNT⁺08]. Nous devons donc faire co-habiter les triplets RDF (qui pourraient être obtenus comme un *mashup* construit grâce aux outils de Bio2RDF) avec des tables traditionnelles comme cela est fait dans certains systèmes de gestion de triplets (*triple stores*) et appliquer les bonnes pratiques de cohabitation telles qu'éditées par des groupes de travail du W3C [BG04]. Afin de concilier le monde du web sémantique et le monde des bases de données, il est possible de s'appuyer sur un moteur Datalog pour permettre, après l'intégration physique des données de sources hétérogènes, l'inférence déductive et la sélection de données pour la fouille ou l'apprentissage, grâce à la récursivité qui manque au langage SQL. Datalog est un sous-ensemble du langage Prolog utilisé comme langage logique d'interrogation de données relationnelles dont le schéma est vu comme un ensemble de prédicats logiques [AHV95]. Des travaux récents proposent un ensemble d'extensions du langage Datalog afin d'optimiser l'interrogation d'ontologies [CGL12].

Un élément important dans les données ouvertes est qu'elles font appel à des ontologies et que ces ontologies soient (ou puissent être) elles-mêmes représentées en RDF. Comme nous

l'avons déjà mentionné, les concepts (ou termes) des ontologies servent à typer les sujets, les prédicats, et les objets [PB13]. Cela permet d'identifier les objets et de repérer des objets équivalents afin de fusionner leurs propriétés lors de l'intégration. Cela permet également de mettre en exergue des conflits ou des incohérences dans les données provenant de plusieurs sources. La provenance des données et des méta-données de qualité sur les sources aideront à la résolution de ces conflits [30, CBL08]. Des services de recherche par le contenu d'ontologies et d'appariement d'ontologies seront certainement des outils précieux pour faire face à la multiplication et à la diversification croissantes des ontologies biomédicales [PHR13, ANS⁺07]. En plus de faciliter l'intégration sémantique des données, les ontologies accessibles par le truchement des DOL constituent un réservoir de connaissances de domaine précieux pour la fouille de données. Ces connaissances *a priori* permettent en positionnant les triplets dans un ou plusieurs référentiels, de favoriser des généralisations et des regroupements tout en évitant une redondance dans les modèles extraits. Cette idée est illustrée par Heiko Paulheim sur un exemple simple [Pau13].

Une fois les données et les ontologies collectées à partir des DOL et d'autres sources plus traditionnelles, la question se pose des méthodes de fouille les plus pertinentes. L'apprentissage par PLI ou les méthodes de fouille de données relationnelles peuvent assez naturellement s'appliquer à l'analyse de triplets tout en prenant en compte la connaissance *a priori*. Les coopérations possibles de ces méthodes avec des méthodes de fouille de graphes reposant sur une représentation simplifiée des données ou de méthodes de fouille de données mono-table me semblent intéressantes à explorer, sur la base d'applications concrètes. Une possibilité est d'isoler par analyse non supervisée ou par échantillonnage des ensembles de données sur lesquels un apprentissage relationnel est faisable, à l'image de l'approche proposée par Grimnes *et al.*, dans laquelle un clustering des données FOAF¹⁷ précède une caractérisation par des règles de logique du premier ordre des différents clusters [GEP04]. Une autre possibilité est de s'appuyer sur les travaux récents de fouille de graphes tels que la recherche efficace de sous-graphes représentatifs d'un ensemble de graphes proposée par Hasan et Zaki [HZ09] afin de réaliser une sélection de prédicats intéressants à explorer par les méthodes relationnelles ou logiques.

5.2.4 Des données patients aux connaissances : vers la médecine personnalisée

L'étude des effets secondaires des molécules que nous avons réalisée se situe au niveau biomoléculaire puisque, partant des effets secondaires notés dans les notices pour les médicaments qui sont sur le marché, l'objectif était de prédire les effets secondaires d'une molécule étant donné ses propriétés et celles de ses cibles [11]. Cela permet, dès la phase de conception du médicament (*drug design*), d'anticiper ses effets indésirables. Ce type d'études, qualifiées de pré-cliniques, sont complémentaires aux études que l'on peut réaliser sur les données cliniques dont on dispose lors des essais cliniques (obligatoires) ou après la mise sur le marché des médicaments lorsque les essais cliniques ont une issue favorable. Les études post-cliniques ont pour but de démontrer puis de s'assurer de la sécurité de l'utilisation des nouveaux médicaments mais aussi de préciser les dosages à adopter. Des études de pharmacovigilance consistent à analyser les données brutes de signalement (par les médecins) des effets secondaires observés sur leurs patients [LMHX12]. Le but est de détecter le plus tôt possible des problèmes liés à la prise d'un médicament ou d'une association de médicaments dans un contexte précis (facteurs démographiques, génétiques,

17. www.foaf-project.org

environnementaux...). Pour ce type d'étude, les chercheurs américains exploitent les données de la base FAERS (*FDA Adverse Event Reporting System*) gérée par la FDA (*Food and Drug Administration*). Dans cette base, les indications et les réactions sont décrites en utilisant l'ontologie MedDRA (utilisée également dans la base de données SIDER). VigiBase est l'équivalent de FAERS au niveau mondial géré par l'OMS (Organisation Mondiale de la Santé).

Un prolongement intéressant de notre étude moléculaire est donc de partir du profil d'effets secondaires prédit par une de nos théories pour de nouvelles molécules (sur le marché) afin :

1. de faire une surveillance focalisée des signalements (reportés dans les ressources telles que FAERS et VigiBase) qui viendraient corroborer la prédiction,
2. d'expliquer dans quel(s) contexte(s) précis ces effets secondaires interviennent.

Cette pharmacovigilance ciblée se justifie pour des effets secondaires délétères (ou des profils comportant de tels effets) et pallie en quelque sorte le fait que nous n'avons pas pris en compte de façon explicite, dans le travail antérieur (Section 3.4.2), l'incertitude dans l'apparition des effets secondaires dans une population. En effet, un attribut fréquence existe dans la base de données SIDER pour qualifier l'association d'une molécule et d'un effet indésirable mais est très rarement renseigné. Nous avons donc choisi d'ignorer cette information, faute de données suffisantes. Ce projet sera l'occasion d'une intégration des données (provenant notamment de FAERS et de VigiBase) et d'un enrichissement (ou contextualisation) de ces données grâce à la collecte de données pertinentes telles que la classification internationale des maladies, la classification des médicaments, celle des indications et des réactions (MedDRA)... Le problème d'apprentissage posé est celui de la recherche de sous-groupes (parmi les patients qui prennent un médicament, quels groupes répondent de façon similaire ?) suivie d'une caractérisation de ces sous-groupes.

Les données patients, qui devraient à terme faire partie du dossier médical personnalisé en France, concernent le suivi médical de patients dans les services hospitaliers. Lorsqu'elles sont disponibles, ces données peuvent être analysées dans le cadre de la médecine personnalisée afin d'améliorer le diagnostic mais aussi la prise en charge c'est à dire la prescription de médicaments ou d'actes médicaux. Parmi ces données, nous pourrions disposer du génotype des patients (CNV, mutations, SNP...), de variables biologiques mesurées tout au long du suivi médical en lien avec des symptômes et des diagnostics (dosages divers, anticorps...), des traitements et/ou des actes médicaux. Le dossier médical d'un patient peut également inclure d'éventuels facteurs environnementaux tels que l'exposition à des facteurs de risque (tabac, polluants, risques professionnels...). Ces données patients, en plus d'être hétérogènes et multi-relationnelles, comportent une dimension temporelle importante à prendre en compte. Elles ont aussi la particularité d'être multi-instances puisque les patients ont des nombres très variables d'épisodes médicaux. Certaines données telles que les données génomiques (SNP, CNV...) sont creuses et nécessitent des modes de représentation et des méthodes statistiques adéquates [YL12] que nous pourrions explorer avec nos collègues bio-statisticiens de l'équipe BIGS de l'IECL (Institut Elie Cartan de Lorraine)¹⁸. En termes de tâches de fouille ou d'apprentissage, la pharmacovigilance ciblée présentée ci-dessus peut s'appliquer sur des données patients sauf que les effets secondaires seront à rechercher dans les données de suivi médical (parmi les symptômes et diagnostics postérieurs à la prise de médicaments).

18. <http://www.inria.fr/equipes/bigs>

Plus généralement, des scénarios d'ECD originaux seront à concevoir afin d'identifier des sous-groupes *homogènes* de patients et de les caractériser en faisant appel si nécessaire à une supervision par des experts en vue de capturer la notion de similarité entre deux patients [SWHE12]. Ainsi, cet environnement d'analyse de données patients pourrait permettre la construction *a posteriori* de cohortes de patients pour diverses analyses pharmacologiques ou en lien avec la santé publique. Enfin, ce projet est en lien avec le projet d'équipe associée Inria SNOWFLAKE coordonné par Adrien Coulet et impliquant un groupe du laboratoire de Russ Altman à l'université de Stanford. Ce groupe amène notamment une composante de fouille de textes indispensable pour l'analyse des parties non structurées des dossiers patients mais aussi pour l'exploitation des données de la littérature scientifique [CCA12].

Dans le cadre du Contrat Plan État-Région (CPER) 2015-2020, un projet multi-disciplinaire est en cours de définition autour de la santé et du vieillissement. Nous participons à ce projet et avons comme objectif d'intervenir assez tôt dans le processus d'élaboration d'essais cliniques ou de constitution de cohortes afin de pouvoir ensuite disposer de données patients de bonne qualité pour effectuer l'extraction de connaissances selon quelques aspects mentionnés dans cette section. Les plateformes (bio-)technologiques et les expertises complémentaires (l'informatique, les bio-statistiques, le bio-medical, la chimie, mais aussi les sciences humaines et sociales) réunies dans ce projet constituent un cadre idéal pour ce faire.

Finalement, les évolutions actuelles autour du *crowd sourcing* nous font espérer une meilleure qualité des données patients d'une part grâce aux efforts qui pourraient être consentis pour l'extraction d'informations à partir des comptes-rendus textuels inclus dans les dossiers médicaux (conformément à l'initiative académique crowd4u¹⁹) et d'autre part par la fourniture directe des données par les patients eux-mêmes à travers des réseaux sociaux spécialisés ou par le biais d'appareils nomades de plus en plus sophistiqués.

5.3 Conclusion

L'activité de recherche que j'ai décrite dans ce document ainsi que les perspectives que j'ai présentées ont été rendues possibles grâce à un environnement stimulant et propice. L'équipe orpailleur est exemplaire en termes d'ouverture vers l'interdisciplinarité puisqu'elle accueille en son sein des chercheurs de disciplines diverses. Grâce à des interactions fortes et privilégiées, nous pouvons y développer des approches symboliques et orientées connaissance. Passer des données biologiques à des connaissances pertinentes nécessite en effet une expertise biologique détenue par les seuls biologistes et du temps pour les non biologistes (dont je fais partie) pour assimiler les concepts importants afin d'arriver à une bonne compréhension des problèmes et proposer des solutions adéquates.

19. <http://crowd4u.org>

Références bibliographiques de l'auteur

- [1] Mehwish Alam, Melisachew Wudage Chekol, Adrien Coulet, Amedeo Napoli, and Malika Smaïl-Tabbone. Lattice Based Data Access (LBDA) : An Approach for Organizing and Accessing Linked Open Data in Biology. In d'Amato et al. [dBSW13].
- [2] Mehwish Alam, Adrien Coulet, Amedeo Napoli, and Malika Smaïl-Tabbone. Formal concept analysis applied to transcriptomic data. In Szathmary and Priss [64], pages 339–344.
- [3] Alexandre Beautrait, Vincent Leroux, Matthieu Chavent, Léo Ghemtio, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Wensheng Cai, Xuegang Shao, Gilles Moreau, Peter Bladon, Jianhua Yao, and Bernard Maigret. Multiple-step virtual screening using VSM-G : overview and validation of fast geometrical matching enrichment. *Journal of Molecular Modeling*, 14(2) :135–148, 2008.
- [4] Sidahmed Benabderrahmane, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Amedeo Napoli, and Olivier Poch. Ontology-based functional classification of genes : evaluation with reference sets and overlap analysis. In Zhongming Zhao, editor, *Proceedings of the 2nd Workshop on Integrative Data Analysis in Systems Biology - IDASB'11*, Atlanta, États-Unis, 2011. IEEE Computer Society.
- [5] Sidahmed Benabderrahmane, Marie-Dominique Devignes, Malika Smaïl-Tabbone, O. Poch, Amedeo Napoli, N. Nguyen N.-H, and W. Raffelsberger. Analyse de données transcriptomiques : Modélisation floue de profils d'expression différentielle et analyse fonctionnelle. In *Actes du XXVIIème congrès Informatique des Organisations et Systèmes d'information et de décision - INFORSID 2009*, pages 413–428, Toulouse, France, 2009.
- [6] Sidahmed Benabderrahmane, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Olivier Poch, Amedeo Napoli, Wolfgang Raffelsberger, Dominique Guenot, Nguyen Hoan, and Eric Guerin. Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets : Application to cancer expression data. In *Actes de la 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances - EGC 2011*, Brest, France, 2011.
- [7] Sidahmed Benabderrahmane, Malika Smaïl-Tabbone, Olivier Poch, Amedeo Napoli, and Marie-Dominique Devignes. IntelliGO : a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11(1) :588, 2010.
- [8] Emmanuel Bresso, Sidahmed Benabderrahmane, Malika Smaïl-Tabbone, Gino Marchetti, Arnaud Sinan Karaboga, Michel Souchet, Amedeo Napoli, and Marie-Dominique Devignes. Use of domain knowledge for dimension reduction : application to mining of drug side effects. In Ana Fred, editor, *Proceedings of the 3rd International Conference on Knowledge Discovery and Information Retrieval-KDIR*, page 8, Paris, France, 2011. INSTICC, SciTePress Digital Library.

- [9] Emmanuel Bresso, Renaud Grisoni, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Formal Concept Analysis for the Interpretation of Relational Learning applied on 3D Protein-Binding Sites. In Ana Fred, editor, *Proceedings of the 4th international conference on Knowledge Discovery and Information Retrieval-KDIR*, page 12 pages, Barcelona, Espagne, 2012. INSTICC, SciTePress Digital Library.
- [10] Emmanuel Bresso, Renaud Grisoni, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. ILP Characterization of 3D Protein-Binding Sites and FCA-Based Interpretation . In A. Fred, K. Dietz, J.L.G.and Liu, and J. Filipe, editors, *Proceedings of the 4th International Joint Conference, IC3K 2012, Barcelona, Spain, October 4-7, 2012. Revised Selected Papers*, volume 415 of *Communications in Computer and Information Science*, pages 84–100. Springer, 2013.
- [11] Emmanuel Bresso, Renaud Grisoni, Gino Marchetti, Arnaud Sinan Karaboga, Michel Souchet, Marie-Dominique Devignes, and Malika Smaïl-Tabbone. Integrative relational machine-learning approach for understanding drug side-effect profiles. *BMC Bioinformatics*, 14 :207, 2013.
- [12] Thibault Caradec, Maude Pupin, Aurélien Vanvlassenbroeck, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Philippe Jacques, and Valérie Leclère. Prediction of monomer isomery in florine : A workflow dedicated to nonribosomal peptide discovery. *PLoS ONE*, 9(1) :e85667, 01 2014.
- [13] Adrien Coulet, Marie-Dominique Devignes, and Malika Smaïl-Tabbone. Extraction de données pharmacogénomiques à partir d'études cliniques : problématique. In *Actes du 2ème atelier sur la "Fouille de données complexes dans un processus d'extraction des connaissances"*, Paris/France, 2005.
- [14] Adrien Coulet, Malika Smaïl-Tabbone, Pascale Benlian, Amedeo Napoli, and Marie-Dominique Devignes. SNP-Converter : an Ontology-Based solution to Reconcile Heterogeneous SNP Descriptions for Pharmacogenomic Studies. In *Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences 2006 - DILS'06*, volume 4075 of *Lecture Notes in Computer Science*, pages 82–93, Hinxton, UK, 2006. Springer.
- [15] Adrien Coulet, Malika Smaïl-Tabbone, Pascale Benlian, Amedeo Napoli, and Marie-Dominique Devignes. SNP-Ontology for semantic integration of genomic variation data (Poster). In *Proceedings of the 14th Annual International Conference on Intelligent Systems for Molecular Biology - ISMB'06*, Fortaleza/Brésil, 2006.
- [16] Adrien Coulet, Malika Smaïl-Tabbone, Pascale Benlian, Amedeo Napoli, and Marie-Dominique Devignes. Ontology-guided Data Preparation for Discovering Genotype-Phenotype Relationships. In *Proceedings of the Network Tools and Applications in Biology : A Semantic Web for Bioinformatics - NETTAB 2007*, Pisa, Italie, 2007.
- [17] Adrien Coulet, Malika Smaïl-Tabbone, Amedeo Napoli, and Marie-Dominique Devignes. Suggested Ontology for Pharmacogenomics (SO-Pharm) : Modular Construction and Preliminary Testing. In *Proceedings of International Workshop on Knowledge Systems in Bioinformatics - KSinBIT'06*, Montpellier, France, 2006.
- [18] Adrien Coulet, Malika Smaïl-Tabbone, Amedeo Napoli, and Marie-Dominique Devignes. Ontology-based knowledge discovery in pharmacogenomics. In Hamid R. Arabnia and Quoc-Nam Tran, editors, *Software Tools and Algorithms for Biological Systems*, Advances in Experimental Medicine and Biology, pages 357–66. Springer, 2011.

- [19] Adrien Coulet, Malika Smaïl-Tabbone, Pascale Benlian, Amedeo Napoli, and Marie-Dominique Devignes. Ontology-guided data preparation for discovering genotype-phenotype relationships. *BMC Bioinformatics*, 9(Suppl 4) :S3, 2008.
- [20] Adrien Coulet, Malika Smaïl-Tabbone, Amedeo Napoli, and Marie-Dominique Devignes. Ontology Refinement through Role Assertion Analysis : Example in Pharmacogenomics. In Franz Baader, Carsten Lutz, and Boris Motik, editors, *Proceedings of the 21st International Workshop on Description Logics - DL2008*, Dresden, Allemagne, 2008.
- [21] Claudia d'Amato, Petr Berka, Vojtech Svátek, and Krzysztof Wecel, editors. *Proceedings of the International Workshop on Data Mining on Linked Data, with Linked Data Mining Challenge collocated with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013)*, Prague, Czech Republic, September 23, 2013, volume 1082 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.
- [22] Jean Devignes, Malika Smaïl-Tabbone, Marie-Dominique Devignes, Alex Hervé, Clotilde Massin, Thomas Lecompte, and Denis Wahl. Profil de positivité initiale et persistance des anticorps antiphospholipides apres un delai de douze semaines. Résumé d'une communication au congrès de la GEHT et de la COMETH à Lille, France, Novembre 2013.
- [23] Marie-Dominique Devignes, Hervé De Palma, Laurent Pierron, Lionel Domenjoud, and Malika Smaïl-Tabbone. User-designed web services to support heterogeneous biological data retrieval. In Angelo Facchiano and Paolo Romano, editors, *Proceedings of the 1st International Workshop on workflows management : new abilities for the biological information overflow - NETTAB 2005*, pages 27–34, Naples, Italy, 2005.
- [24] Marie-Dominique Devignes, Philippe Franiatte, Nizar Messai, Emmanuel Bresso, Amedeo Napoli, and Malika Smaïl-Tabbone. BioRegistry : Automatic extraction of metadata for biological database retrieval and discovery. *International Journal of Metadata Semantics and Ontologies*, 5(3) :184–193, 2010.
- [25] Marie-Dominique Devignes, Philippe Franiatte, Nizar Messai, Amedeo Napoli, and Malika Smaïl-Tabbone. BioRegistry : Automatic Extraction of Metadata for Biological Database Retrieval and Discovery. In *Proceedings of the 1st international workshop on Ressource Discovery (RED), Joint to iiWAS 2008*, Linz, Autriche, 2008.
- [26] Marie-Dominique Devignes, Nizar Messai, Amedeo Napoli, Shazia Osman, and Malika Smaïl-Tabbone. Intelligent access to genomic sources on the web. In *Proceedings of the W3C Workshop on Semantic Web for Life Sciences*, Cambridge, Massachusetts USA, 2004.
- [27] Marie-Dominique Devignes, Yvan Norsa, Malika Smaïl-Tabbone, Philippe Collet, Lionel Domenjoud, and Michel Dauça. A Generic Solution for Automated Collecting and Integration of Biological Data from Web Sources (Poster). In *Proceedings of the European Conference on Computational Biology - ECCB'03*, page 2, Paris, France, 2003.
- [28] Marie-Dominique Devignes, André Schaaff, and Malika Smaïl. Collecte et intégration de données biologiques hétérogènes sur le web : application dans le domaine de la cartographie du génome humain. *Ingénierie des Systèmes d'Information*, 7(1-2) :45–61, 2002.
- [29] Marie-Dominique Devignes, Benabderrahmane Sidahmed, Malika Smaïl-Tabbone, Napoli Amedeo, and Poch Olivier. Functional classification of genes using semantic distance and fuzzy clustering approach : Evaluation with reference sets and overlap analysis. *international Journal of Computational Biology and Drug Design. Special Issue on : "Systems Biology Approaches in Biological and Biomedical Research"*, 5(3/4) :245–260, 2012.

- [30] Marie-Dominique Devignes and Malika Smaïl-Tabbone. Integration of biological data from web resources : management of multiple answers through metadata retrieval. In *Proceedings of the 12th International Conference on Intelligent Systems for Molecular Biology - 3rd European Conference on Computational Biology - ISMB-ECCB 2004*, page 3, Glasgow, Scotland, Royaume-Uni, 2004.
- [31] Marie-Dominique Devignes and Malika Smaïl-Tabbone. Maîtriser les ressources numériques : biologie "in silico". In Magali Roux, editor, *Biologie L'ère numérique*, pages 189–222. CNRS Editions, 2009.
- [32] Marie-Dominique Devignes and Malika Smaïl-Tabbone. Workshop PC Chairs' Message. Web Data Integration for Mining in the Life Sciences (WebDIM4LS). In Mathias Weske, Mohand-Said Hacid, and Claude Godart, editors, *Proceedings of the International Workshops on Web Information Systems Engineering – WISE 2007 Workshops*, volume 4832 of *Lecture Notes in Computer Science*, pages 3–4, Nancy, France, 2007. Springer.
- [33] Gérald Duffing and Malika Smaïl. A Novel Approach for Accessing Partially Indexed Image Corpora. In *Proceedings of the 4th International Conference on Visual Information Systems - VISUAL'2000*, page 13 p, Lyon, France, 2000.
- [34] Gérald Duffing and Malika Smaïl. Organising and Searching Partially Indexed Image Databases. In F. Crestani, M. Girolami, and C. J van Rijsbergen, editors, *Proceedings of the 24th BCS-IRSG European Colloquium on Information Retrieval Research - ECIR 2002*, volume 2291 of *Lecture Notes in Computer Science*, pages 22–40, Glasgow, Scotland, UK, 2002. Springer Verlag.
- [35] Leo Ghemtio, Emmanuel Bresso, Michel Souchet, Bernard Maigret, Malika Smaïl-Tabbone, and Marie-Dominique Devignes. Model-driven data integration for mining protein-ligand and protein-protein interactions in a drug design context. In *Actes des Journées Ouvertes Biologie Informatique Mathématiques - JOBIM 2008*, page 2p., Lille, France, 2008.
- [36] Leo Ghemtio, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Michel Souchet, Vincent Leroux, and Bernard Maigret. Comparison of three preprocessing filters efficiency in virtual screening : identification of new putative LXRbeta regulators as a test case. *Journal of chemical information and modeling*, 50(5) :701–715, 2010.
- [37] Leo Ghemtio, Malika Smaïl-Tabbone, Appolinaire Djikeng, Marie-Dominique Devignes, Lionel Keminse, Patricia Kelbert, Joseph Fokam, Bernard Maigret, and Odile Ouwe-Missi-Oukem-Boyer. HIV-PDI : A Protein-Drug Interaction Resource for Structural Analyses of HIV Drug Resistance : 1. Concepts and Associated Database. *Journal of Health & Medical Informatics*, 2(1) :1000104, 2011.
- [38] Léo Ghemtio, Malika Smaïl-Tabbone, Marie-Dominique Devignes, Michel Souchet, and Bernard Maigret. A KDD Approach for Designing Filtering Strategies to Improve Virtual Screening. In Ana Fred, editor, *Proceedings of the 1st International Conference on Knowledge Discovery and Information Retrieval - KDIR*, Madeira, Portugal, 2009. INSTIC.
- [39] Anisah Ghoorah, Marie-Dominique Devignes, Malika Smaïl-Tabbone, and David Ritchie. Spatial clustering of protein binding sites for template based protein docking. *Bioinformatics*, 27(20) :2820–2827, 2011.
- [40] Renaud Grisoni, Marie-Dominique Bresso, Emmanuel and Devignes, and Malika M. Smaïl-Tabbone. Méthodologie et outils pour l'extraction de connaissances par Programmation Logique Inductive (PLI) (Poster). In *Actes de la 13ème Conférence*

Francophone sur l'Extraction et la Gestion des Connaissances- EGC 2013, Toulouse, France, 2013.

- [41] Rachid Hafiane, Malika Smaïl-Tabbone, Marie-Dominique Devignes, and Salvatore Tabbone. Clustering optimal de gènes fondé sur une mesure de similarité sémantique. In *Actes 10ème édition de la Conférence en Recherche d'Information et Applications - CORIA 2013*, page 15 p, Neuchâtel, Suisse, 2013.
- [42] Nathalie Leblond-Bourget, Gérard Guédon, Sophie Payot, Yann Aubert, Malika Smaïl-Tabbone, and Marie-Dominique Devignes. Icefinder : Knowledge-based identification of integrative conjugative elements (ice) in newly sequenced genomes using coding sequence signatures (poster). In *Actes des Journées Ouvertes Biologie Informatique Mathématiques (JOBIM 2013)*, page 2p., Toulouse, France, 2013.
- [43] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Classification et interrogation de sources de données biologiques. *Revue des Nouvelles Technologies de l'Information RNTI*, pages 43–47, 2005.
- [44] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Querying a Bioinformatic Data Sources Registry with Concept Lattices. In Marie-Laure Mugnier Frithjof Dau and Gerd Stumme, editors, *Proceedings of the 3rd International Conference on Conceptual Structures - ICCS 2005*, volume 3596 of *LNAI*, pages 323–336, Kassel, Allemagne, 2005. Springer Berlin / Heidelberg.
- [45] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. BR-Explorer : A sound and complete FCA-based retrieval algorithm (Poster). In *Proceedings of the 4th International Conference on Formal Concept Analysis - ICFCA 2006*, Dresden/Germany, 2006.
- [46] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. BR-Explorer : An FCA-based algorithm for Information Retrieval. In *Fourth International Conference On Concept Lattices and Their Applications - CLA 2006*, Hammamet/Tunisia, 2006.
- [47] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques BioRegistry. *Ingénierie des Systèmes d'Information (ISI)*, 11(1) :39–60, 2006.
- [48] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Correction et complétude d'un algorithme de recherche d'information par treillis de concepts. *Revue des Nouvelles Technologies de l'Information RNTI*, 2007.
- [49] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Traitement d'attributs inter-dépendants pour la recherche d'information par treillis. In *Actes de la conférence francophone Ingénierie des Connaissances - IC 2007*, pages 109–120, Grenoble, France, 2007. Cépaduès éditions.
- [50] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Many-Valued Concept Lattices for Conceptual Clustering and Information Retrieval. In Malik Ghallab, editor, *Proceedings of the 18th European Conference in Artificial Intelligence - ECAI 2008*, pages 127–131, Patras, Grèce, 2008. IOS Press.
- [51] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Connaissances de domaine et treillis de concepts pour l'exploration progressive de données complexes. In Sylvie DESPRES, editor, *Acte des 21èmes Journées francophones d'Ingénierie des Connaissances*, pages 233–244, Nîmes, France, 2010. Ecole des Mines d'Alès.

- [52] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Using Domain Knowledge to Guide Lattice-based Complex Data Exploration. In Rudi Studer Helder Coelho and Michael Wooldridge, editors, *Proceedings of the 19th European Conference on Artificial Intelligence - ECAI 2010*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 847–852, Lisbon, Portugal, 2010. IOS press.
- [53] Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, and Malika Smaïl-Tabbone. Extending Attribute Dependencies for Lattice-Based Querying and Navigation. In Peter W. Eklund and Ollivier Haemmerlé, editors, *Proceedings of the 16th International Conference on Conceptual Structures - ICCS 2008*, volume 5113 of *Lecture Notes in Computer Science*, pages 189–202, Toulouse, France, 2008. Springer.
- [54] Nizar Messai, Marie-Dominique Devignes, Malika Smaïl-Tabbone, and Amedeo Napoli. Treillis de concepts et ontologies pour l'interrogation d'un annuaire de sources de données biologiques (BioRegistry). In *Actes du XXIIIème congrès Informatique des Organisations et Systèmes d'information et de décision - INFORSID 2005*, Grenoble/France, 2005.
- [55] Birama Ndiaye, Emmanuel Bresso, Malika Smaïl-Tabbone, Michel Souchet, and Marie-Dominique Devignes. Modim : Model-driven data integration for mining (poster). In *Actes des Journées Ouvertes Biologie Informatique Mathématiques (JOBIM 2011)*, page 2p., Paris, France, 2011.
- [56] Gabin Personeni, Simon Daget, Céline Bonnet, Philippe Jonveaux, Marie-Dominique Devignes, Malika Smaïl-Tabbone, and Adrien Coulet. Model-driven integration of linked open data for mining : a case study with genes responsible for intellectual disability. In *Proceedings of the 10th International Conference on Data Integration in the Life Sciences 2014 - DILS 2014*, volume to appear of *Lecture Notes in Bioinformatics*, Hinxton, UK, 2014. Springer.
- [57] Maude Pupin, Malika Smaïl-Tabbone, Philippe Jacques, Marie-Dominique Devignes, and Valérie Leclère. NRPS toolbox for the discovery of new nonribosomal peptides and synthetases. In François Coste et Denis Tagu, editor, *Actes des Journées Ouvertes en Biologie, l'Informatique et les mathématiques (JOBIM) 2012*, pages 89–93, Rennes, France, 2012.
- [58] Brigitte Simonnot and Malika Smaïl. Model for interactive retrieval of videos and still images. In Kingsley C. Nwosu P. Bruce Berra and Bhavani Thuraisingham, editors, *Multimedia database systems*, pages 278–317. Kluwer, 1996.
- [59] Malika Smaïl. Case-Based Information Retrieval. In S. Wess, Althoff K., and M. Richter, editors, *Proceedings of the 1st European Workshop on Topics in Case-Based Reasoning, Selected Papers EWCBR 1993*, volume 837 of *Lecture Notes in Computer Science*, pages 404–413. Springer, 1993.
- [60] Malika Smaïl. Vers des systèmes évolutifs de recherche d'information : un état de l'art. *Revue Technique et Science Informatiques - TSI*, 17(10) :1193–1222, 1998.
- [61] Malika Smaïl. Recherche de régularités dans une mémoire de sessions de recherche d'information documentaire. In *Actes du XVIIème congrès Informatique des Organisations et Systèmes d'information et de décision - INFORSID 1999*, page 19 p, La Garde, Toulon, 1999.
- [62] Malika Smaïl and Marion Crehange. Case-Based Reasoning Meets Information Retrieval. In J-L. Funck-Brentano and F. Seitz, editors, *Proceedings of the 4th International Conference on Computer-Assisted Information Retrieval - RIAO 1994*, pages 172–185, New York, USA, 1994.

- [63] Malika Smaïl-Tabbone, Shazia Osman, Nizar Messai, Amedeo Napoli, and Marie-Dominique Devignes. BioRegistry : a structured metadata repository for bioinformatic databases. In M.R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, and I. Fisher, editors, *Proceedings of the 1st International Symposium on Computational Life Science - CompLife 2005*, volume 3695 of *Lecture Notes in Bioinformatics*, pages 46–56, Konstanz, Germany, 2005. Springer.
- [64] Laszlo Szathmary and Uta Priss, editors. *Proceedings of the 9th International Conference on Concept Lattices and Their Applications, Fuengirola (Málaga), Spain, October 11-14, 2012*, volume 972 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.
- [65] Saliha Yilmaz, Philippe Jonveaux, Cedric Bicep, Laurent Pierron, Malika Smaïl-Tabbone, and Marie-Dominique Devignes. Gene-Disease Relationship Discovery based on Model-driven Data Integration and Database View Definition. *Bioinformatics*, 25(2) :230 – 236, 2009.

Références bibliographiques

- [AAD⁺96] Sameet Agarwal, Rakesh Agrawal, Prasad M. Deshpande, Ashish Gupta, Jeffrey F. Naughton, Raghu Ramakrishnan, and Sunita Sarawagi. On the computation of multidimensional aggregates. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 506–521. Morgan Kaufmann, 1996.
- [ACMG13] Alessandro Antonucci, Giorgio Corani, Denis Deratani Mauá, and Sandra Gabaglio. An ensemble of bayesian networks for multilabel classification. In Rossi [Ros13].
- [AF10a] Tarek Abudawood and Peter A. Flach. The advantages of seed examples in first-order multi-class subgroup discovery. In Coelho et al. [CSW10], pages 1113–1114.
- [AF10b] Tarek Abudawood and Peter A. Flach. Learning Multi-class Theories in ILP. In Frasconi and Lisi [FL11], pages 6–13.
- [AHV95] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
- [AKF⁺13] Harith Alani, Lalana Kagal, Achille Fokoue, Paul T. Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors. *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*. Springer, 2013.
- [ANS⁺07] Harith Alani, Natasha Fridman Noy, Nigam Shah, Nigel Shadbolt, and Mark A. Musen. Searching ontologies based on content : experiments in the biomedical domain. In Sleeman and Barker [SB07], pages 55–62.
- [AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB’94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994.
- [AS96] Setsuo Arikawa and Arun Sharma, editors. *Algorithmic Learning Theory, 7th International Workshop, ALT ’96, Sydney, Australia, October 23-25, 1996, Proceedings*, volume 1160 of *Lecture Notes in Computer Science*. Springer, 1996.
- [ASPS98] Rakesh Agrawal, Paul E. Stolorz, and Gregory Piatetsky-Shapiro, editors. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York City, New York, USA, August 27-31, 1998*. AAAI Press, 1998.

- [AVB01] Frédéric Achard, Guy Vaysseix, and Emmanuel Barillot. Xml, bioinformatics and data integration. *Bioinformatics*, 17(1) :115–125, 2001.
- [AW10] Charu C. Aggarwal and Haixun Wang. *Managing and Mining Graph Data*. Springer Publishing Company, Incorporated, 1st edition, 2010.
- [BB01] P. Baldi and S. Brunak. *Bioinformatics : The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.
- [BBB⁺08] Fadi Badra, Rokia Bendaoud, Rim Bentebibel, Pierre-Antoine Champin, Julien Cojan, Amélie Cordier, Sylvie Desprès, Stéphanie Jean-Daubias, Jean Lieber, Thomas Meilender, Alain Mille, Emmanuel Nauer, Amedeo Napoli, and Yannick Toussaint. Taaable : Text mining, ontology engineering, and hierarchical classification for textual case-based cooking. In Schaaf [Sch08], pages 219–228.
- [BBDF07] Sarah Cohen Boulakia, Olivier Biton, Susan B. Davidson, and Christine Froidevaux. BioGuideSRS : Querying Multiple Sources with a User-centric Perspective. *Bioinformatics*, 23(10) :1301–1303, 2007.
- [BBF08] Amos Bairoch, Sarah Cohen Boulakia, and Christine Froidevaux, editors. *Data Integration in the Life Sciences, 5th International Workshop, DILS 2008, Evry, France, June 25-27, 2008. Proceedings*, volume 5109 of *Lecture Notes in Computer Science*. Springer, 2008.
- [BBJN96] Franz Baader, Martin Buchheit, Manfred A. Jeusfeld, and Werner Nutt, editors. *Knowledge Representation Meets Databases, Proceedings of the 3rd Workshop KRDB’96, Budapest, Hungary, August 13, 1996*, volume 4 of *CEUR Workshop Proceedings*. CEUR-WS.org, 1996.
- [BCD⁺09] M. R. Berthold, N. Cebon, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. KNIME - the Konstanz information miner : version 2.0 and beyond. *SIGKDD Explorations*, 11(1) :26–31, November 2009.
- [BCF⁺10] Hendrik Blockeel, Toon Calders, Elisa Fromont, Bart Goethals, Adriana Prado, and Céline Robardet. *Inductive Databases and Constraint-Based Data Mining*, chapter A Practical Comparative Study Of Data Mining Query Languages, pages 59–77. Computer Science. Springer, December 2010.
- [BCF⁺12] Hendrik Blockeel, Toon Calders, Élisabeth Fromont, Bart Goethals, Adriana Prado, and Céline Robardet. An inductive database system based on virtual mining views. *Data Min. Knowl. Discov.*, 24(1) :247–287, 2012.
- [BCT06] Zohra Bellahsene, Robbie Coenmans, and John Tranier. Matérialisation de vues dans les entrepôts de données. une approche dynamique. *Ingénierie des Systèmes d’Information*, 11(6) :33–53, 2006.
- [BD01] Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Williams College, Williamstown, MA, USA, June 28 - July 1, 2001. Morgan Kaufmann, 2001.
- [BDD⁺02] Hendrik Blockeel, Luc Dehaspe, Bart Demoen, Gerda Janssens, Jan Ramon, and Henk Vandecasteele. Improving the Efficiency of Inductive Logic Programming Through the Use of Query Packs. *J. Artif. Intell. Res. (JAIR)*, 16 :135–166, 2002.

- [Ben09] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1) :1–127, 2009.
- [Ben11] Sidahmed Benabderrahmane. *Prise en compte des connaissances du domaine dans l'analyse transcriptomique : Similarité sémantique, classification fonctionnelle et profils flous. Application au cancer colorectal*. PhD Thesis in Computer Science, Université Henri Poincaré - Nancy I, 2011.
- [BF06] L. Barbosa and J. Freire. Combining classifiers to identify online databases. In *16th international conference on World Wide Web*, pages 431–440, 2006.
- [BFOS84] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- [BG04] Dave Beckett and Jan Grant. SWAD-Europe Deliverable 10.2 : Mapping Semantic Web Data with RDBMSes, 2004.
- [BHBL09] 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.
- [BM05] J-F. Boulicaut and C. Masson. Data mining query languages. In O. Maimon and L. Rokach, editors, *The Data Mining and Knowledge Discovery Handbook*, pages 715–727. Springer, 2005.
- [BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets : Generalizing association rules to correlations. In Peckham [Pec97], pages 265–276.
- [BMS07] Michael R. Berthold, Katharina Morik, and Arno Siebes, editors. *Parallel Universes and Local Patterns*, volume 07181 of *Dagstuhl Seminar Proceedings*, 2007.
- [BNT⁺08] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF : Towards a mashup to build bioinformatics knowledge systems. *J. of Biomedical Informatics*, 41(5) :706–716, October 2008.
- [BOS⁺05] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alexander J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. In *ISMB (Supplement of Bioinformatics)* [DBL05], pages 47–56.
- [BPT⁺00] Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, and Lotfi Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In Lloyd et al. [LDF⁺00], pages 972–986.
- [Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1) :5–32, 2001.
- [Bre13] Emmanuel Bresso. *Organisation et exploitation des connaissances sur les réseaux d'interactions biomoléculaires pour l'étude de l'étiologie des maladies génétiques et la caractérisation des effets secondaires de principes actifs*. PhD thesis, University de Lorraine, 2013.
- [BRM05] Jean-François Boulicaut, Luc De Raedt, and Heikki Mannila, editors. *Constraint-Based Mining and Inductive Databases, European Workshop on Inductive Databases and Constraint Based Mining, Hinterzarten, Germany, March 11-13, 2004, Revised Selected Papers*, volume 3848 of *Lecture Notes in Computer Science*. Springer, 2005.
- [BS03] Hendrik Blockeel and Michèle Sebag. Scalability and efficiency in multi-relational data mining. *SIGKDD Explorations*, 5(1) :17–30, 2003.

- [BTN08] Rokia Bendaoud, Yannick Toussaint, and Amedeo Napoli. Pactole : A methodology and a system for semi-automatically enriching an ontology from a collection of texts. In Peter W. Eklund and Ollivier Haemmerlé, editors, *International Conference on Conceptual Structures*, volume 5113 of *Lecture Notes in Computer Science*, pages 203–216. Springer, 2008.
- [BYYO11] Michelle D. Brazas, David S. Yim, Joseph Tadashi Yamada, and B. F. Francis Ouellette. The 2011 Bioinformatics links directory update : more resources, tools and databases and features to empower the bioinformatics community. *Nucleic Acids Research*, 39(Web-Server-Issue) :3–7, 2011.
- [CBL08] James Cheney, Peter Buneman, and Bertram Ludäscher. Report on the principles of provenance workshop. *SIGMOD Record*, 37(1) :62–65, 2008.
- [CCA12] Adrien Coulet, K. Bretonnel Cohen, and Russ B. Altman. The state of the art in text mining and natural language processing for pharmacogenomics. *Journal of Biomedical Informatics*, 45(5) :825–826, 2012.
- [CCP⁺13] Philipp Cimiano, Óscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors. *The Semantic Web : Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, volume 7882 of *Lecture Notes in Computer Science*. Springer, 2013.
- [CCTAD13] Alison Callahan, Jose Cruz-Toledo, Peter Ansell, and Michel Dumontier. Bio2RDF Release 2 : Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In Cimiano et al. [CCP⁺13], pages 200–212.
- [CGL12] Andrea Calì, Georg Gottlob, and Thomas Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. *Journal of Web Semantics*, 14 :57–83, 2012.
- [CHST04] Philipp Cimiano, Andreas Hotho, Gerd Stumme, and Julien Tane. Conceptual knowledge processing with formal concept analysis and ontologies. In Peter Eklund, editor, *Concept Lattices*, volume 2961 of *Lecture Notes in Computer Science*, pages 189–207. Springer, Berlin/Heidelberg, 2004.
- [CKS04] Rui Camacho, Ross D. King, and Ashwin Srinivasan, editors. *Inductive Logic Programming, 14th International Conference, ILP 2004, Porto, Portugal, September 6-8, 2004, Proceedings*, volume 3194 of *Lecture Notes in Computer Science*. Springer, 2004.
- [CLM⁺09] Amélie Cordier, Jean Lieber, Pascal Molli, Emmanuel Nauer, Hala Skaf-Molli, and Yannick Toussaint. Wikitaaable : A semantic wiki as a blackboard for a textual case-base reasoning system. In Lange et al. [LSSMV09].
- [CM98] Gregory F. Cooper and Serafín Moral, editors. *UAI '98 : Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, University of Wisconsin Business School, Madison, Wisconsin, USA, July 24-26, 1998*. Morgan Kaufmann, 1998.
- [CM07] Vladimir S. Cherkassky and Filip Mulier. *Learning from Data : Concepts, Theory, and Methods*. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition, 2007.
- [CMA05] Nicola Cannata, Emanuela Merelli, and Russ B. Altman. Time to organize the bioinformatics resourceome. *PLoS Computational Biology*, 1(7), 2005.

- [Con08] BioMoby Consortium. Interoperability with moby 1.0- it's better than sharing your toothbrush! *Briefings in Bioinformatics*, 9(3) :220–231, 2008.
- [Con10] The Gene Ontology Consortium. The gene ontology in 2010 : extensions and refinements. *Nucleic Acids Research*, 38 :D331–D335, 2010.
- [CRB04] Toon Calders, Christophe Rigotti, and Jean-François Boulicaut. A survey on condensed representations for frequent sets. In Boulicaut et al. [BRM05], pages 64–80.
- [CSW10] Helder Coelho, Rudi Studer, and Michael Wooldridge, editors. *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*, volume 215 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2010.
- [CSZ06] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
- [DBdCD⁺05] Jesse Davis, Elizabeth S. Burnside, Inês de Castro Dutra, David Page, and Vítor Santos Costa. An integrated approach to learning bayesian networks of rules. In Gama et al. [GCB⁺05], pages 84–95.
- [DBL97] *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 23-29, 1997, 2 Volumes*. Morgan Kaufmann, 1997.
- [DBL05] *Proceedings Thirteenth International Conference on Intelligent Systems for Molecular Biology 2005, Detroit, MI, USA, 25-29 June 2005*, 2005.
- [dBSW13] Claudia d’Amato, Petr Berka, Vojtech Svátek, and Krzysztof Wecel, editors. *Proceedings of the International Workshop on Data Mining on Linked Data, with Linked Data Mining Challenge collocated with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013), Prague, Czech Republic, September 23, 2013*, volume 1082 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.
- [dCDPCS02] Inês de Castro Dutra, David Page, Vítor Santos Costa, and Jude W. Shavlik. An empirical evaluation of bagging in inductive logic programming. In Matwin and Sammut [MS03], pages 48–65.
- [Del88] Jean-Paul Delahaye. *Outils logiques pour l’intelligence artificielle*. Eyrolles, 1988.
- [DF99] Saso Džeroski and Peter A. Flach, editors. *Inductive Logic Programming, 9th International Workshop, ILP-99, Bled, Slovenia, June 24-27, 1999, Proceedings*, volume 1634 of *Lecture Notes in Computer Science*. Springer, 1999.
- [DKK05] Mukund Deshpande, Michihiro Kuramochi, and George Karypis. Mining chemical compounds. In Jason Tsong-Li Wang, Mohammed Javeed Zaki, Hannu Toivonen, and Dennis Shasha, editors, *Data Mining in Bioinformatics*, pages 189–215. Springer, 2005.
- [DL01] Saso Džeroski and Nada Lavrač. *Relational Data Mining*. Springer-Verlag New York, Inc., 2001.
- [DLBL⁺12] Valmi Dufour-Lussier, Florence Le Ber, Jean Lieber, Thomas Meilender, and Emmanuel Nauer. Semi-automatic annotation process for procedural texts : An application on cooking recipes. *CoRR*, abs/1209.5663, 2012.

- [DOS⁺07] Jesse Davis, Irene M. Ong, Jan Struyf, Elizabeth S. Burnside, David Page, and Vítor Santos Costa. Change of representation for statistical relational learning. In Veloso [Vel07], pages 2719–2726.
- [DS04] Frank DiMaio and Jude W. Shavlik. Learning an approximation to inductive logic programming clause evaluation. In Camacho et al. [CKS04], pages 80–97.
- [DS07] Saso Džeroski and Jan Struyf, editors. *Knowledge Discovery in Inductive Databases, 5th International Workshop, KDID 2006, Berlin, Germany, September 18, 2006, Revised Selected and Invited Papers*, volume 4747 of *Lecture Notes in Computer Science*. Springer, 2007.
- [DT01] Luc Dehaspe and Hannu Toivonen. Discovery of relational association rules. In Sašo Džeroski and Nada Lavrač, editors, *Relational Data Mining*, pages 189–208. Springer-Verlag New York, Inc., New York, NY, USA, 2001.
- [DTK98] Luc Dehaspe, Hannu Toivonen, and Ross D. King. Finding frequent substructures in chemical compounds. In Agrawal et al. [ASPS98], pages 30–36.
- [DW03] M. Dekkers and S. Weibel. State of the dublin core metadata initiative. *D-Lib Magazine*, 9(4), 2003.
- [DWCH12] Krzysztof Dembczynski, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2) :5–45, 2012.
- [EK03] D.P. Enot and D.K. King. Application of inductive logic programming to structure-based drug design, 2003.
- [EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis et al. [SHF96], pages 226–231.
- [ELLS11] Brian Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis*. Wiley Series in probability and statistics. John Wiley & Sons, 2011.
- [ERC⁺13] Elias Egho, Chedy Raïssi, Toon Calders, Thomas Bourquard, Nicolas Jay, and Amedeo Napoli. Vers une mesure de similarité pour les séquences complexes. In Vrain et al. [VPS13], pages 335–340.
- [ESBB98] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25) :14863–14868, 1998.
- [FBS06] Élisabeth Fromont, Hendrik Blockeel, and Jan Struyf. Integrating decision tree learning into inductive databases. In Džeroski and Struyf [DS07], pages 81–96.
- [FGW02] Usama Fayyad, Georges G. Grinstein, and Andreas Wierse, editors. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.
- [FL11] Paolo Frasconi and Francesca A. Lisi, editors. *Inductive Logic Programming - 20th International Conference, ILP 2010, Florence, Italy, June 27-30, 2010. Revised Papers*, volume 6489 of *Lecture Notes in Computer Science*. Springer, 2011.
- [FMPS98] Paul W. Finn, Stephen Muggleton, David Page, and Ashwin Srinivasan. Pharmacophore discovery using the inductive logic programming system progol. *Machine Learning*, 30(2-3) :241–270, 1998.

- [FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17 :37–54, 1996.
- [FQC05] Élisabeth Fromont, René Quiniou, and Marie-Odile Cordier. Learning rules from multisource data for cardiac monitoring. In Miksch et al. [MHK05], pages 484–493.
- [FR09] S. Ferré and S. Rudolph, editors. *Formal Concept Analysis, 7th International Conference, ICFCA 2009, Darmstadt, Germany, May 21-24, 2009, Proceedings*, volume 5548 of *Lecture Notes in Computer Science*. Springer, 2009.
- [FSG13] Xosé M. Fernández-Suarez and M. Y. Galperin. The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 41(Database-Issue) :1–7, 2013.
- [Gö3] Thomas Gärtner. A survey of kernels for structured data. *SIGKDD Explorations Newsletters*, 5(1) :49–58, July 2003.
- [GAV98] Alexander Gammerman, Katy S. Azoury, and Vladimir Vapnik. Learning by transduction. In Cooper and Moral [CM98], pages 148–155.
- [GCB⁺05] João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo, editors. *Machine Learning : ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, volume 3720 of *Lecture Notes in Computer Science*. Springer, 2005.
- [GEP04] Gunnar Aastrand Grimnes, Peter Edwards, and Alun D. Preece. Learning meta-descriptions of the foaf network. In McIlraith et al. [MPvH04], pages 152–165.
- [GFKT01] Lise Getoor, Nir Friedman, Daphne Koller, and Benjamin Taskar. Learning probabilistic models of relational structure. In Brodley and Danyluk [BD01], pages 170–177.
- [GG09] Jean-Gabriel Ganascia and Pierre Gançarski, editors. *Extraction et gestion des connaissances (EGC’2009), Actes, Strasbourg, 27 au 30 janvier 2009*, volume RNTI-E-15 of *Revue des Nouvelles Technologies de l’Information*. Cépaduès-Éditions, 2009.
- [GGMW03] Prasanna Ganesan, Hector Garcia-Molina, and Jennifer Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1) :64–93, 2003.
- [GH07] Fabrice Guillet and Howard J. Hamilton, editors. *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*. Springer, 2007.
- [GMB⁺05] Emilie Guérin, Gwenaëlle Marquet, Anita Burgun, Olivier Loréal, Laure Berti-Equille, Ulf Leser, and Fouzia Moussouni. Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW. In Ludäscher and Raschid [LR05], pages 158–174.
- [GMMT07] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. *TKDD*, 1(3), 2007.
- [GPE05] Asunción Gómez-Pérez and Jérôme Euzenat, editors. *The Semantic Web : Research and Applications, Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29 - June 1, 2005, Proceedings*, volume 3532 of *Lecture Notes in Computer Science*. Springer, 2005.

- [GSSB00a] Attilio Giordana, Lorenza Saitta, Michèle Sebag, and Marco Botta. Analyzing relational learning in the phase transition framework. In Langley [Lan00], pages 311–318.
- [GSSB00b] Attilio Giordana, Lorenza Saitta, Michèle Sebag, and Marco Botta. Can relational learning scale up? In Ras and Ohsuga [RO00], pages 31–39.
- [GTDD07] Pierre Geurts, Nizar Touleimat, Marie Dutreix, and Florence d’Alché Buc. Inferring biological networks with output kernel trees. *BMC Bioinformatics*, 8(S-2), 2007.
- [GW99] B. Ganter and R. Wille. *Formal Concept Analysis*. Springer, Berlin, 1999.
- [GWS03] Carole Goble, Christopher Wroe, and Robert Stevens. The mygrid project : services, architecture and demonstrator. In *All Hands Meeting*, pages 595–603, September 2003.
- [HB11] Tom Heath and Christian Bizer. *Linked Data : Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.
- [HBS⁺09] Syed Haider, Benoit Ballester, Damian Smedley, Junjun Zhang, Peter Rice, and Arek Kasprzyk. BioMart Central Portal – unified access to biological data. *Nucleic Acids Research*, 37(suppl 2) :W23–W27, 2009.
- [HHNV13] Mohamed Rouane Hacene, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. Relational concept analysis : mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell.*, 67(1) :81–108, 2013.
- [HK01] J. Han and M. Kamber. *Data Mining : Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2001.
- [HMP97] David Heckerman, Heikki Mannila, and Daryl Pregibon, editors. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, California, USA, August 14-17, 1997*. AAAI Press, 1997.
- [Hor03] Tamás Horváth, editor. *Inductive Logic Programming : 13th International Conference, ILP 2003, Szeged, Hungary, September 29-October 1, 2003, Proceedings*, volume 2835 of *Lecture Notes in Computer Science*. Springer, 2003.
- [HPT01] Mohand-Said Hacid, Jean-Marc Petit, and Farouk Toumani. Representing and reasoning on database conceptual schemas. *Knowledge and Information Systems*, 3(1) :52–80, 2001.
- [HPT⁺02] Lynette Hirschman, Jong C. Park, Junichi Tsujii, Limsoon Wong, and Cathy H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12) :1553–1561, 2002.
- [HT98] Laura M. Haas and Ashutosh Tiwary, editors. *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*. ACM Press, 1998.
- [HZ09] Mohammad Al Hasan and Mohammed J. Zaki. Output space sampling for graph patterns. *PVLDB (Proceedings of the VLDB Endowment)*, 2(1) :730–741, 2009.
- [IM96] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *Communications of ACM*, 39(11) :58–64, 1996.

- [KCFS08] A. Knobbe, Bruno Crémilleux, J. Fürnkranz, and M. Scholz. From local patterns to global models : The lego approach to data mining. In *International Workshop From Local Patterns to Global Models co-located with ECML/PKDD'08*, pages 1–16, Antwerp, Belgium, September 2008.
- [KCL⁺10] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars J. Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(1), 2010.
- [Kei02] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1) :1–8, January 2002.
- [KEV05] Martin Krallinger, Ramon Alonso-Allende A. Erhardt, and Alfonso Valencia. Text-mining approaches in molecular biology and biomedicine. *Drug discovery today*, 10(6) :439–445, 2005.
- [KKS⁺04] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. EnsMart : A Generic System for Fast and Flexible Access to Biological Data. *Genome Research*, 14(1) :160–169, January 2004.
- [KLW08] Peter D. Karp, Thomas J. Lee, and Valerie Wagner. BioWarehouse : Relational Integration of Eleven Bioinformatics Databases and Formats. In Bairoch et al. [BBF08], pages 5–7.
- [KO02] Sergei O. Kuznetsov and Sergei Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental and Theoretical Artificial Intelligence*, 14 :189–216, 2002.
- [KR05] L. Kaufman and J. Rousseeuw, P. *Finding groups in data - An introduction to Cluster Analysis*. Wiley Series in probability and statistics. Wiley, 2005.
- [KRZ⁺03] Mark-A. Krogel, Simon Rawles, Filip Zelezný, Peter A. Flach, Nada Lavrač, and Stefan Wrobel. Comparative evaluation of approaches to propositionalization. In Horváth [Hor03], pages 197–214.
- [KSS95] Ross D. King, Michael J. E. Sternberg, and Ashwin Srinivasan. Relating Chemical Activity to Structure : An Examination of ILP Successes. *New Generation Comput.*, 13(3 and 4) :411–433, 1995.
- [KSS⁺07] Craig Knox, Savita Shrivastava, Paul Stothard, Roman Eisner, and David S. Wishart. BioSpider : a web server for automating metabolome annotations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 145–156, 2007.
- [Kuz07] S.O. Kuznetsov. On stability of a Formal Concept. *Ann. Math. Artif. Intell.*, 49(1-4) :101–115, 2007.
- [Lan00] Pat Langley, editor. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, Stanford University, Stanford, CA, USA, June 29 - July 2, 2000. Morgan Kaufmann, 2000.
- [LAWG05] Phillip W. Lord, Pinar Alper, Chris Wroe, and Carole A. Goble. Feta : A light-weight architecture for user oriented semantic service discovery. In Gómez-Pérez and Euzenat [GPE05], pages 17–31.
- [LBE03] Zoé Lacroix, Omar Boucelma, and Mehdi Essid. The biological integration system. In *Proceedings of the 5th ACM international workshop on Web*

- information and data management*, WIDM '03, pages 45–49, New York, NY, USA, 2003. ACM.
- [LDB96] N. Lavrač, S. Džeroski, and I. Bratko. Handling imperfect data in inductive logic programming. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 48–64. IOS, 1996.
- [LDF⁺00] John W. Lloyd, Verónica Dahl, Ulrich Furbach, Manfred Kerber, Kung-Kiu Lau, Catuscia Palamidessi, Luís Moniz Pereira, Yehoshua Sagiv, and Peter J. Stuckey, editors. *Computational Logic - CL 2000, First International Conference, London, UK, 24-28 July, 2000, Proceedings*, volume 1861 of *Lecture Notes in Computer Science*. Springer, 2000.
- [Len02] M. Lenzerini. Data integration : A theoretical perspective. In *PODS'02, 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.
- [LFZ99] Nada Lavrač, Peter A. Flach, and Blaz Zupan. Rule Evaluation Measures : a Unifying View. In Džeroski and Flach [DF99], pages 174–185.
- [Llo93] John Wylie Lloyd. *Foundations of Logic Programming*. Springer-Verlag, Secaucus, NJ, USA, 2nd edition, 1993.
- [LMHX12] Mei Liu, Michael E. Matheny, Yong Hu, and Hua Xu. Data mining methodologies for pharmacovigilance. *SIGKDD Explorations*, 14(1) :35–42, 2012.
- [LN05] Ulf Leser and Felix Naumann. (almost) hands-off information integration for the life sciences. In *Second biennial Conference on Innovative Data systems Research*, pages 131–143, California, 2005.
- [LNST06] Jean Lieber, Amedeo Napoli, Laszlo Szathmary, and Yannick Toussaint. First elements on knowledge discovery guided by domain knowledge (kddk). In Sadok Ben Yahia, Engelbert Mephu Nguifo, and Radim Belohlavek, editors, *Concept Lattices and Their Applications, Fourth International Conference, CLA 2006, Selected Papers*, volume 4923 of *Lecture Notes in Computer Science*, pages 22–41, Tunisia, 2006. Springer.
- [LR05] Bertram Ludäscher and Louiqa Raschid, editors. *Proceedings of the 2nd workshop on Data Integration in the Life Sciences, DILS 2005, CA, USA, July 20-22, 2005*, volume 3615 of *Lecture Notes in Computer Science*. Springer, 2005.
- [LSSMV09] Christoph Lange, Sebastian Schaffert, Hala Skaf-Molli, and Max Völkel, editors. *4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Greece, June 1st, 2009. Proceedings*, volume 464 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.
- [LTP07] Stéphane Lallich, Olivier Teytaud, and Elie Prudhomme. Association rule interestingness : Measure and statistical validation. In Guillet and Hamilton [GH07], pages 251–275.
- [MBR⁺04] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox : functional analysis of gene datasets based on Gene Ontology. *Genome Biology*, 5(12) :R101+, 2004.
- [MHK05] Silvia Miksch, Jim Hunter, and Elpida T. Keravnou, editors. *Artificial Intelligence in Medicine, 10th Conference on Artificial Intelligence in Medicine*,

- AIME 2005, Aberdeen, UK, July 23-27, 2005, Proceedings*, volume 3581 of *Lecture Notes in Computer Science*. Springer, 2005.
- [Mit82] Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2) :203–226, 1982.
- [MNMHEB12] Ali Masoudi-Nejad, Alireza Meshkin, Behzad Haji-Eghrari, and Gholamreza Bidkhor. Candidate gene prioritization. *Molecular genetics and genomics : MGG*, 287(9) :679–698, 2012.
- [MPvH04] Sheila A. McIlraith, Dimitris Plexousakis, and Frank van Harmelen, editors. *The Semantic Web - ISWC 2004 : Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004. Proceedings*, volume 3298 of *Lecture Notes in Computer Science*. Springer, 2004.
- [MR94] S. Muggleton and L. De Raedt. Inductive logic programming : Theory and methods. *The Journal of Logic Programming*, 19(20) :629–679, 1994.
- [MS03] Stan Matwin and Claude Sammut, editors. *Inductive Logic Programming, 12th International Conference, ILP 2002, Sydney, Australia, July 9-11, 2002. Revised Papers*, volume 2583 of *Lecture Notes in Computer Science*. Springer, 2003.
- [MSKS98] S. Muggleton, A. Srinivasan, R. D. King, and M. J. E. Sternberg. Biochemical knowledge discovery using inductive logic programming. In Setsuo Arikawa and Hiroshi Motoda, editors, *Discovery Science, First International Conference, DS '98, Fukuoka, Japan, December 14-16, 1998, Proceedings*, volume 1532 of *Lecture Notes in Computer Science*, pages 326–341. Springer, 1998.
- [MSRO⁺10] Paolo Missier, Stian Soiland-Reyes, Stuart Owen, Wei Tan, Aleksandra Nenadic, Ian Dunlop, Alan Williams, Thomas Oinn, and Carole Goble. Taverna, reloaded. In M. Gertz, T. Hey, and B. Ludaescher, editors, *SSDBM 2010*, Heidelberg, Germany, 2010.
- [MSTH05] Peter Mork, Ron Shaker, and Peter Tarczy-Hornoch. The multiple roles of ontologies in the biomediator data integration system. In *Proceedings of the Second international conference on Data Integration in the Life Sciences*, pages 96–104, Berlin, Heidelberg, 2005. Springer-Verlag.
- [MT97] Heikki Mannila and Hannu Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3) :241–258, 1997.
- [Mug91] S. Muggleton. Inductive logic programming. *New Generation Computing*, 8(4) :295–318, 1991.
- [Mug95] Stephen Muggleton. Inverse entailment and progol. *New Generation Comput.*, 13(3&4) :245–286, 1995.
- [Mug99] Stephen Muggleton. Scientific knowledge discovery using inductive logic programming. *Commun. ACM*, 42(11) :42–46, 1999.
- [NH08] Thanh Phuong Nguyen and Tu Bao Ho. An integrative domain-based approach to predicting protein-protein interactions. *J. Bioinformatics and Computational Biology*, 6(6) :1115–1132, 2008.
- [NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering : Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.

- [NLHP98] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Han, and Alex Pang. Exploratory mining and pruning optimizations of constrained association rules. In Haas and Tiwary [HT98], pages 13–24.
- [NLW09] Petra Kralj Novak, Nada Lavrač, and Geoffrey I. Webb. Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10 :377–403, 2009.
- [ODC⁺08] Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovanni Tummarello. Sindice.com : a document-oriented lookup index for open linked data. *IJMSO*, 3(1) :37–52, 2008.
- [Pau13] Heiko Paulheim. Exploiting linked open data as background knowledge in data mining. In d’Amato et al. [dBSW13].
- [PB13] Heiko Paulheim and Christian Bizer. Type Inference on Noisy RDF Data. In Alani et al. [AKF⁺13], pages 510–525.
- [PC03] D. Page and M. Craven. Biological applications of multi-relational data mining. *SIGKDD Explorations*, 5(1) :69–79, 2003.
- [Pec97] Joan Peckham, editor. *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*. ACM Press, 1997.
- [PF97] Foster J. Provost and Tom Fawcett. Analysis and visualization of classifier performance : Comparison under imprecise class and cost distributions. In Heckerman et al. [HMP97], pages 43–48.
- [PFF⁺09] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), 2009.
- [PHL04] Jian Pei, Jiawei Han, and Laks V. S. Lakshmanan. Pushing convertible constraints in frequent itemset mining. *Data Mining and Knowledge Discovery*, 8(3) :227–252, 2004.
- [PHR13] Heiko Paulheim, Sven Hertling, and Dominique Ritze. Towards evaluating interactive ontology matching tools. In Cimiano et al. [CCP⁺13], pages 31–45.
- [PLC⁺08] Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. Linking data to ontologies. In *J. Data Semantics* [Spa08], pages 133–173.
- [Plo70] G. Plotkin. A note on inductive generalization. *Machine Intelligence*, 5, 1970.
- [Qui96] J. Ross Quinlan. Boosting first-order learning. In Arikawa and Sharma [AS96], pages 143–155.
- [Rae97] Luc De Raedt. Logical settings for concept-learning. *Artificial Intelligence*, 95(1) :187–201, 1997.
- [Rae08] Luc De Raedt. *Logical and Relational Learning*. Cognitive Technologies. Springer, 2008.
- [RBF⁺02] M.C. Rousset, A. Bidault, C. Froidevaux, H. Gagliardi, F. Goasdoué, C. Reynaud, and B. Safar. Construction de médiateurs pour intégrer des sources d’information multiples et hétérogènes : le projet PICSEL. *Revue I3*, 2(1) :9–59, 2002.

- [Res07] V.A. Reston. MedDRA Maintenance and Support Services Organization. Introductory Guide, MedDRA Version 10.1. International Federation of Pharmaceutical Manufacturers and Associations, 2007.
- [RK03] Luc De Raedt and Kristian Kersting. Probabilistic logic learning. *SIGKDD Explorations*, 5(1) :31–48, 2003.
- [RLKB03] Volker Roth, Julian Laub, Motoaki Kawanabe, and Joachim M. Buhmann. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12) :1540–1551, 2003.
- [RO00] Zbigniew W. Ras and Setsuo Ohsuga, editors. *Foundations of Intelligent Systems, 12th International Symposium, ISMIS 2000, Charlotte, NC, USA, October 11-14, 2000, Proceedings*, volume 1932 of *Lecture Notes in Computer Science*. Springer, 2000.
- [Ros13] Francesca Rossi, editor. *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*. IJCAI/AAAI, 2013.
- [Rou94] Céline Rouveirol. Flattening and saturation : Two representation changes for generalization. *Machine Learning*, 14(1) :219–232, 1994.
- [SB07] Derek H. Sleeman and Ken Barker, editors. *Proceedings of the 4th International Conference on Knowledge Capture (K-CAP 2007), October 28-31, 2007, Whistler, BC, Canada*. ACM, 2007.
- [SBB⁺00] Robert Stevens, Patricia Baker, Sean Bechhofer, Gary, Alex Jacoby, Norman W. Paton, Carole Goble, and Andy Brass. Tambis : Transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2) :184–186, 2000.
- [Sch08] Martin Schaaf, editor. *ECCBR 2008, The 9th European Conference on Case-Based Reasoning, Trier, Germany, September 1-4, 2008, Workshop Proceedings*, 2008.
- [SFC⁺06] H. Sun, H. Fang, T. Chen, R. Perkins, and W. Tong. GOFFA : Gene Ontology For Functional Analysis - A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data. *BMC Bioinformatics*, 7 Suppl 2 :23, 2006. Journal article BMC Bioinformatics. 2006 Sep 26 ;7 Suppl 2 :S23.
- [SHF96] Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*. AAAI Press, 1996.
- [SMKV11] Timos K. Sellis, Renée J. Miller, Anastasios Kementsietsidis, and Yannis Velegrakis, editors. *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011*. ACM, 2011.
- [Sou06] Arnaud Soulet. *Un cadre générique de découverte de motifs sous contraintes fondées sur des primitives*. PhD thesis, University de Caen Basse-Normandie, 2006.
- [Spa08] Stefano Spaccapietra, editor. *Journal on Data Semantics X*, volume 4900 of *Lecture Notes in Computer Science*. Springer, 2008.
- [SR97] Michèle Sebag and Céline Rouveirol. Tractable induction and classification in first order logic via stochastic matching. In *IJCAI (2)* [DBL97], pages 888–893.

- [Sri07] Ashwin Srinivasan. The aleph manual. available at <http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/aleph/>, 2007.
- [SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [SSW⁺08] Damian Smedley, Morris A. Swertz, Katy Wolstencroft, Glenn Proctor, Michael Zouberakis, Jonathan B. L. Bard, John M. Hancock, and Paul N. Schofield. Solutions for data integration in functional genomics : a critical assessment and case study. *Briefings in Bioinformatics*, 9(6) :532–544, 2008.
- [SWHE12] Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi. Supervised patient similarity measure of heterogeneous patient records. *SIGKDD Explorations*, 14(1) :16–24, 2012.
- [Sza06] L. Szathmary. *Symbolic Data Mining Methods with the Coron Platform*. PhD Thesis in Computer Science, University Henri Poincaré – Nancy 1, France, Nov 2006.
- [TASM08] Kazuhisa Tsunoyama, Ata Amini, Michael J. E. Sternberg, and Stephen H. Muggleton. Scaffold Hopping in Drug Discovery Using Inductive Logic Programming. *J. Chem. Inf. Model.*, 48(5) :949–957, May 2008.
- [TKS02] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 32–41, New York, NY, USA, 2002. ACM.
- [TMS01] M Turcotte, S Muggleton, and M Sternberg. Automated discovery of structural signatures of protein fold and function. *Journal of Molecular Biology*, 306(3) :591–605, 2001.
- [Ull00] Jeffrey D. Ullman. Information integration using logical views. *Theor. Comput. Sci.*, 239(2) :189–210, 2000.
- [Vap99] Vladimir Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5) :988–999, 1999.
- [Vel07] Manuela M. Veloso, editor. *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, 2007.
- [VPS13] Christel Vrain, André Péninou, and Florence Sèdes, editors. *Extraction et gestion des connaissances (EGC'2013), Actes, 29 janvier - 01 février 2013, Toulouse, France*, volume RNTI-E-24 of *Revue des Nouvelles Technologies de l'Information*. Hermann-Éditions, 2013.
- [Web07] G. I. Webb. Discovering Significant Patterns. *Machine Learning*, 68(1) :1–33, 2007.
- [WF05] I. Witten and E. Frank. *Data Mining : Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.
- [Wie92] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3) :38–49, 1992.
- [WM03] Takashi Washio and Hiroshi Motoda. State of the art of graph-based data mining. *SIGKDD Explor. Newsl.*, 5(1) :59–68, July 2003.
- [WPCB⁺03] Jason Weston, Fernando Pérez-Cruz, Olivier Bousquet, Olivier Chapelle, André Elisseeff, and Bernhard Schölkopf. Feature selection and transduction for

- prediction of molecular bioactivity for drug design. *Bioinformatics*, 19(6) :764–771, 2003.
- [WW07] Silke Wagner and Dorothea Wagner. Comparing Clusterings : An Overview . Technical Report 2006-04, Universität Karlsruhe (TH), 2007.
- [WZTS05] J.T.L. Wang, M.J. Zaki, H. Toivonen, and D. Shasha. *Data Mining in Bioinformatics*. Advanced Information and Knowledge Processing. Springer, 2005.
- [YHYY06] Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu. Efficient classification across multiple database relations : A crossmine approach. *IEEE Trans. Knowl. Data Eng.*, 18(6) :770–783, 2006.
- [YL12] Jieping Ye and Jun Liu. Sparse methods for biomedical data. *SIGKDD Explorations*, 14(1) :4–15, 2012.
- [YR12] Shipeng Yu and Bharat Rao. Introduction to the special section on clinical data mining. *SIGKDD Explorations*, 14(1) :1–3, 2012.
- [ZLAE02] Evgeni Zdobnov, Rodrigo Lopez, Rolf Apweiler, and Thure Etzold. The EBI SRS server – recent developments. *Bioinformatics*, 18 :2002, 2002.
- [ZPZ⁺13] Fei Zhu, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, and Bairong Shen. Biomedical text mining and its applications in cancer research. *Journal of biomedical informatics*, 46(2) :200–211, 2013.

Dossier de présentation

Malika SMAIL épouse TABBONE

Née en 1965 en Algérie

Mariée, 4 enfants nés en 1996, 1999, et des jumeaux en 2001

Courriel : Malika.Smail@loria.fr

Téléphone : 03 83 59 20 65

Adresse professionnelle

LORIA UMR 7503,

Campus Scientifique BP 239,

54 506 Vandœuvre-lès-Nancy

1 Diplômes universitaires

2014 Habilitation à Diriger des Recherches Titre du document : Contributions à l'extraction de connaissances à partir de données biologiques. Devant le jury : Christine Froidevaux, Céline Rouveirol, Julie Thompson, Anne Boyer, Marie-Dominique Devignes, Bruno Crémilleux

1994 Doctorat en Informatique obtenu à l'Université Henri Poincaré, Nancy 1.

Titre de la thèse : Raisonnement à base de cas pour une recherche évolutive d'information.

1990 DEA en Informatique de l'Université Henri Poincaré, Nancy 1.

1989 Diplôme d'ingénieur en informatique obtenu à l'université de Tizi-Ouzou en Algérie.

2 Expérience professionnelle et statut actuel

1995- Maître de Conférences hors-classe en Informatique à l'Université de Lorraine, Département d'informatique et membre du laboratoire LORIA (UMR 7503). Titulaire de la PES depuis 2013.

2007-2009 Accueil en délégation à l'INRIA (CRI Nancy-Grand Est) dans l'équipe-projet Orpailleur.

1993-1995 Attaché Temporaire d'Enseignement et de Recherche (ATER) à l'Université de Nancy 2 puis à l'Université Henri Poincaré, Nancy 1.

3 Résumé de mon activité de recherche

Parcours

Je résume ici mon parcours de recherche de la thèse de doctorat à aujourd'hui. Ma thèse a porté sur l'application du raisonnement à partir de cas pour rendre évolutif un système de recherche d'images indexées par des mot-clés [59, 62, 58, 60, 61]. Nous avons ensuite, dans le

cadre de la thèse de Gérald Duffing que j'ai co-encadrée avec Marion Créhange, proposé une approche thématico-visuelle pour l'organisation et l'interrogation interactive d'une collection d'images. Une double description des images selon un axe thématique et un axe visuel s'appuie sur l'indexation de l'image à l'aide mots-clés, d'une part, et sur sa caractérisation visuelle grâce à des indices tels que la couleur ou la texture, d'autre part. Dans ce contexte, la collection d'images peut être organisée sous forme de deux structures hiérarchiques - les dendrogrammes - faisant apparaître les similitudes existant entre les images au niveau thématique pour l'une et au niveau visuel pour l'autre. Cette organisation autorise la définition d'opérations de manipulation et de coopération entre ces structures, avec l'objectif de mettre à la disposition de la recherche d'images un ensemble de mécanismes mettant en évidence cette interdépendance : il devient dès lors possible de retrouver des images faisant référence à une thématique précise et répondant à des critères visuels donnés [33, 34].

Le départ de Marion Créhange à la retraite et l'extinction de l'équipe EXPRIM dont elle était responsable ont coïncidé avec l'arrivée de Marie-Dominique Devignes, chargée de recherche CNRS en Sciences de la Vie, qui a décidé de demander son rattachement au LORIA suite son installation en Lorraine. Marie-Dominique et moi avons décidé de travailler ensemble sur des problèmes d'intégration et de recherche d'information puis d'extraction de connaissances à partir de sources de données biologiques hétérogènes. J'ai ainsi réorienté ma recherche vers des problématiques assez différentes de mes activités antérieures et cela a nécessité que je me familiarise avec ces nouvelles thématiques de la section 27 et que je m'initie à ce nouveau domaine qu'est la biologie.

Marie-Dominique et moi avons, depuis lors, travaillé en étroite interaction et avons progressé l'une en informatique et l'autre en biologie tout d'abord dans l'équipe Langue et Dialogue dirigée par Jean-Marie Pierrel. Nous avons assez rapidement évolué vers l'extraction de connaissances à partir de bases de données (biologiques) qui constitue une suite logique à l'intégration de ces données. C'est ainsi que nous avons rejoint l'équipe Orpailleur dirigée par Amedeo Napoli en 2003 où nous avons co-encadré plusieurs thèses et stages de DEA puis de Master relevant de l'école doctorale IAEM (Informatique Automatique Electronique et Mathématiques) mais aussi BioSE (Biologie Santé et Environnement) ou CPM (Chimie et Physique Moléculaires).

Problématique de recherche

Après ma thèse et quelques années de recherche dans le domaine de la recherche d'images, je me suis intéressée, avec M-D. Devignes, à l'intégration de données à partir de sources de données biologiques hétérogènes. En effet, l'étude du monde du vivant a donné lieu à une multitude de banques de données dont la croissance en taille et en complexité est exponentielle [31]. Un des défis majeurs de l'ère post-génomique (après le séquençage complet de nombreux génomes dont le génome humain) a tout d'abord consisté à faciliter la recherche de ces données dispersées dans des sources hétérogènes auxquelles le web a banalisé l'accès. Nous avons tout d'abord abordé le problème de l'intégration de données hétérogènes de façon pragmatique en privilégiant le point de vue de l'utilisateur. Nous avons ainsi proposé et expérimenté, à travers le logiciel Xcollect et le cadre du mémoire CNAM d'André Schaaff, une approche générique pour automatiser la collecte de données selon un scénario (ou workflow) défini par l'utilisateur et faisant appel à des sources de données bien identifiées [28, 27]. Nous avons également étudié le problème des réponses homologues (réponses à une même question retournées par plusieurs sources de données) et proposé la prise en compte de certains critères de qualité d'une source (tels la fréquence de mise à jour, le mode de validation des contenus) afin d'aider le biologiste à trier ces divers résultats [30]. Notre approche a été appliquée à la recherche de données cartographiques associées à des gènes

d'intérêt puis à la recherche de gènes candidats pour des maladies rares telles que le syndrome d'Aicardi [65].

Nous nous sommes ensuite intéressés à la collecte et l'organisation des méta-données relatives aux sources de données afin de faciliter la sélection, voire la découverte de sources de données pertinentes pour un besoin précis. Nous avons ainsi construit un annuaire appelé BioRegistry (<http://bioregistry.loria.fr>) organisant les méta-données de façon structurée et faisant appel à des ontologies de domaine et proposé des mécanismes de recherche de sources de données [63, 25]. Dans le cadre de la thèse de Nizar Messai, nous avons appliqué l'analyse de concepts formels (FCA) pour organiser les sources du BioRegistry selon les propriétés (ou méta-données) qu'elles partagent et permettre aux utilisateurs de naviguer dans le treillis obtenu [50, 53, 52].

En plus de recherches ponctuelles de données dans les sources de données (identifiées grâce au BioRegistry ou déjà connues), il est nécessaire, pour de nombreux problèmes biologiques, d'intégrer puis de fouiller de gros volumes de données en vue d'en extraire de nouveaux éléments de connaissance. Ceci nous amène naturellement à nous intéresser à l'extraction de connaissances à partir de données (ECD ou KDD pour *Knowledge Discovery from Databases*) incluant, en plus du processus de fouille lui-même, l'intégration en amont des données dans une base de données et l'interprétation en aval des résultats de la fouille, précédant une éventuelle itération du processus.

Nous avons implémenté dans un logiciel, MODIM, le principe de l'intégration de données dirigée par un modèle relationnel afin de faciliter aussi bien la préparation, la fouille que l'interprétation des résultats de la fouille [55] grâce au contrat IJD (Ingénieur Jeune Diplômé) Inria de Birama Ndiaye. Ce principe avait émergé à l'occasion de la recherche de gènes candidats pour expliquer une maladie rare dans le cadre de la thèse de Saliha Yilmaz [65]. Le logiciel MODIM a ensuite été utilisé dans le cadre de la thèse de Léo Ghemtio pour collecter et intégrer dans une base les données sur les interactions protéine-ligand pour le criblage virtuel et la conception de nouveaux médicaments [38, 36]. De nombreuses autres utilisations ont suivi [57, 8, 42].

Nous étudions dans l'équipe Orpailleur un raffinement du processus de KDD par la prise en compte autant que possible de connaissances du domaine aux différentes étapes du processus [52, 4, 18]. La première étude que nous avons menée était dans le cadre de la pharmacogénomique et de la thèse d'Adrien Coulet, la pharmacogénomique consistant à analyser les relations entre profil génétique, absorption de médicaments, et phénotype observé (réaction aux médicaments) en utilisant une formalisation en logique de descriptions des données et connaissances [16, 19, 18]. La thèse de Sidahmed Benabderrahmane sur l'analyse des données du transcriptome liées au cancer colo-rectal nous a ensuite amenés à définir une mesure de similarité entre entités biologiques, IntelliGO. Cette mesure est inspirée des mesures utilisées en recherche d'information prenant en compte des relations sémantiques entre les éléments du vocabulaire d'indexation, Gene Ontology qui se présente comme un graphe acyclique orienté [5, 7, 4]. Cette mesure a été afin de réduire la dimension des données relatives aux effets secondaires des médicaments [8].

Les deux années de ma délégation à l'INRIA m'ont permis d'amorcer une thématique de recherche propre : la fouille de données relationnelles pour l'analyse de données biologiques complexes comportant des aspects relationnels et structurels importants. La complexité des données biologiques m'a amenée à m'intéresser à des méthodes de fouille capables de considérer des données dans leur format relationnel et qui ne nécessitent pas leur transformation préalable en matrice, voire en matrice binaire objets×attributs. C'est ainsi que nous avons utilisé la Programmation Logique Inductive (PLI) pour la caractérisation et la prédiction de sites spécifiques d'interactions de protéines (stage de master d'Emmanuel Bresso en 2009) [10]. La PLI a ensuite été appliquée pour l'analyse de données issues d'une base de connaissances pharmacogénomiques en incluant des connaissances du domaine (stage de Master d'Andreea Orosanu en 2010), à l'analyse de profils d'effets secondaires de médicaments et des gènes responsables de maladies [11, 56].

Je m'intéresse également à diverses combinaisons de méthodes de fouille notamment pour l'aide à l'interprétation de résultats de fouille [9] ou pour la construction de modèles globaux selon l'approche LeGo [KCFS08]. Nous avons également, grâce au contrat IJD de Renaud Grisoni, pu implémenter la chaîne du processus d'ECD intégrant un programme de PLI et de FCA dans l'environnement KNIME [40].

Dans sa thèse en cours, Mehwish Alam étudie l'utilisation d'une extension de la FCA aux données relationnelles, connue sous le nom de RCA, pour l'analyse de données biologiques relatives aux gènes dérégulés chez des patients souffrant de cancers pour lesquels aucun traitement n'est proposé à l'heure actuelle (données fournies par des partenaires du consortium Bio-Intelligence) [2]. Elle s'intéresse également à la construction de treillis de concepts formels pour faciliter l'organisation et l'accès aux données ouvertes [1].

Par ailleurs, les méthodes de fouille dites symboliques sont souvent opposées aux méthodes dites statistiques. La combinaison de ces deux types de méthodes fait l'objet précisément du projet ComSSyCo que j'anime dans le cadre du CPER 2007-2013 en collaboration avec des collègues de l'équipe BIGS de Bio-Statistiques de l'Institut Elie Cartan (Aurélien Muller-Gueudin) sur des problèmes d'analyses de données de cohortes fournies par des collègues de l'unité INSERM 954 (Nutrition - génétique et exposition aux risques environnementaux).

En résumé, le fil conducteur de mon activité de recherche est l'intégration de données et l'extraction de connaissances à partir de données relationnelles en biologie et dans le domaine bio-médical.

4 Activités d'encadrement

Encadrement de thèses

09/2012- Encadrement avec Amedeo Napoli de la thèse de Mehwish ALAM

Titre : Combinaisons de méthodes pour l'extraction de connaissances guidée par les connaissances du domaine.

11/2005-10/2008 Encadrement avec MD. Devignes de la thèse d'Adrien COULET

Titre : Intégration et extraction de connaissances à partir de données cliniques et génétiques.

10/2004-03/2009 Encadrement avec MD. Devignes et A. Napoli de la thèse de Nizar MESSAI

Titre : Extraction de connaissances et Web sémantique : application à la recherche et l'interrogation de ressources génomiques sur le Web.

09/2006-05/2010 Encadrement avec Bernard Maigret de la thèse de Léo GHEMTIO inscrit à l'école doctorale *Chimie et Physique Moléculaires*

Titre : Simulation numérique et approche orientée connaissance pour la découverte de nouvelles molécules thérapeutiques.

09/1996-12/1999 Encadrement avec Marion Créhange (Prof. Emérite, Université de Nancy 2) de la thèse de Gérard DUFFING (actuellement enseignant à l'Institut Commercial de Nancy) Titre : Apports mutuels de la vision par ordinateur et de la recherche interactive d'images.

Devenir des docteurs

- Gérard Duffing est enseignant-chercheur à l'ICN (Institut Commercial de Nancy).
- Adrien Coulet est maître de conférences à l'ESIAL devenue Télécom Nancy (école d'ingénieurs, informatique, Université de Lorraine).

- Nizar Messai est maître de conférences à l'université de Tours.
- Léo Ghemtio est post-doctorant à Helsinki.

Encadrement d'autres travaux

- 02/2010-09/2010** Encadrement avec MD Devignes du stage de Master 2 Informatique d'Andreea Orosanu
 Titre : Fouille de données relationnelles au sein d'une base de connaissances pharmacogénomique.
- 02/2009-09/2009** Encadrement avec MD Devignes du stage de Master 2 Sciences de la Vie et de la Santé spécialité *Génomique et informatique* d'Emmanuel Bresso (poursuite en thèse CIFRE)
 Titre : Prise en compte de la structure 3D des sites de phosphorylation en vue de leur caractérisation explicite par deux méthodes de fouille de données symboliques.
- 08/2007-08/2008** Encadrement avec MD-Devignes du mémoire CNAM (1 année) de Philippe Franiatte (retour en entreprise)
 Titre : Conception et la réalisation d'un système de recherche de ressources biologiques.
- 02/2004-06/2004** Encadrement avec MD Devignes du stage de M2 Informatique de Nizar Messai (poursuite en thèse)
 Titre : Treillis de Galois et ontologies de domaine pour la classification et la recherche de sources de données génomiques.
- 01/1999-12/2000** Encadrement avec MD Devignes du mémoire CNAM (1 année) d'André Schaauff (actuellement ingénieur de recherche au Centre de Données astronomiques de Strasbourg)
 Titre : Conception et réalisation d'un système de recherche d'informations sur la cartographie du génome humain.

Valorisation de la recherche

- 2009-2012** Je suis associée au dépôt APP (Agence pour la Protection des Programmes) de trois logiciels, MODIM : logiciel de collecte de données dirigée par un modèle de données (2009) ; IntelliGO : mesure de similarité sémantique entre objets décrits par des concepts structurés en graphes acycliques orientés ou DAG) (2010), WAFBI : nœuds pour la plateforme KNIME pour la fouille de données relationnelles à partir de bases de données (2013).
- 2007-2009** Membre du comité de pilotage du Centre de Compétences et de Transfert G-BioModel (Génomique et modélisation des bio-molécules) qui a donné lieu à la société Harmonic Pharma dirigée par Michel Souchet provenant de l'industrie pharmaceutique²⁰. La société bénéficie du concours scientifique de mes deux collègues M-D. Devignes et B. Maigret.
- 2005-2008** Participation au projet EUREKA GenNet. Les partenaires industriels sont Kika Medical (société française) et Phenosystems (société Belge). Co-encadrement d'un doctorant en contrat CIFRE, Adrien COULET. J'ai fait un exposé sur cette expérience à une journée de sensibilisation des chercheurs à l'importance du transfert industriel organisée par le service RIV (Relations Internationales et Valorisation) en juin 2011.

20. <http://www.harmonicpharma.com>

Animation de la recherche

- Membre du comité d'organisation de la conférence européenne ECCB'14 (*European Conference on Computational Biology*) à Strasbourg en septembre 2014. Co-chair pour les Posters.
- Membre du comité d'organisation de la *Semaine du Document Numérique et de la Recherche d'Information* (SDNRI'14) qui a réuni les deux conférences francophones CORIA et CIFED à Nancy en mars 2014.
- Co-chair, en octobre 2007 d'un workshop associé à la conférence internationale WISE'07 (Web Information Systems Engineering) intitulé *Web Data Integration and Mining for Life Sciences*. Les actes de l'ensemble des workshops de la conférence ont été publiés dans un volume LNCS (4832).
- Co-organisation avec Mario Albrecht (Max Plack Institut Informatik, Saarbruck), M-D. Devignes et Dave Ritchie (DR INRIA), du premier workshop Sarr-Lor-Lux intitulé "Computational, Structural and Medical Approaches for Systems Biology", LORIA, Nancy, 14-15 décembre 2009.
- Participation au thème MBI, Modélisation des Biomolécules et de leurs Interactions, du CPER MISN (2007-2013). J'y ai coordonné l'opération-projet IP3L (Interactions Protéine-Protéine et Protéine-Ligand) de 2008 à 2010. Cela a porté sur la conception et la mise en place d'une base de données IP3L sur les interactions protéine-protéine et protéine-ligand afin de faciliter aussi bien la fouille de données que l'interprétation des résultats de la fouille. Nous avons peuplé la base de données à l'aide des données sur les interactions protéine-ligand pour le criblage virtuel et la conception de nouveaux médicaments. Diverses méthodes de fouille de données symboliques ont été utilisées sur ces données (motifs fréquents, règles d'association, arbres de décision). La PLI a également été appliquée aux données d'interaction protéine-protéine.
- 2004-2007 : Membre du comité scientifique de l'ACI soutenue par l'ANR (programme IMP-BIO) intitulée ISIBio (Intégration des Systèmes d'Information en Biologie) et coordonnée par M-D. Devignes. Ce groupe de travail national a regroupé plusieurs équipes de laboratoires français (LRI/Orsay, LIRMM/Montpellier, INRA/Jouy, CIRAD/Montpellier, INSERM/Rennes) et avait pour objectif de développer une animation scientifique dans le domaine des systèmes d'information en biologie afin de renforcer les collaborations existantes, d'initier de nouvelles interactions et d'accroître la visibilité internationale de la communauté française dans cette thématique. Les thématiques abordées étaient celles de l'interopérabilité des ressources hétérogènes, les architectures de médiation, la représentation et gestion des connaissances, la construction et le partage d'ontologies. Du côté de la biologie, quelques domaines d'application ont été ciblés dont l'analyse du transcriptome. Sur ses 3 années de fonctionnement, le groupe de travail ISIBio a organisé 3 séminaires thématiques (en 2005 et 2006), deux journées satellites de la conférence JO-BIM (2005 et 2006) et le workshop satellite de la conférence internationale WISE en 2007. Nous avons à chacune de ces occasions invité au moins un conférencier étranger spécialisé dans la thématique (Phillip Lord, Arek Kazprzyk, Susan Davidson, Yves Moreau, Guido Vetere, Emmanuella Merelli). Le rapport final de l'ACI ISIBio est accessible à www.loria.fr/~devignes/ISIBio-ACI_IMPBio-RapportDeFinDeProjet.pdf.

2010-2013 : J'ai été responsable du projet ComSSyCo dans le cadre du CPER 2007-2013, thème MBI. Ce projet a porté sur la comparaison et la combinaison de méthodes de fouille statistiques et symboliques pour l'analyse de données de cohortes en vue de l'extraction de connaissances. Deux partenaires sont associés à ce projet : une équipe de l'IECN (Institut Elie Cartan de Nancy) spécialisée dans les bio-statistiques et une équipe de l'INSERM (U954) dirigée par le Professeur J-L Guéant et qui chargée de fournir des données de cohorte ainsi que des besoins d'analyse.

Co-organisation de trois journées scientifiques autour de la bioinformatique dans le cadre du CPER MISN (25 nov. 2004, 16 déc. 2005, 25-26 oct. 2012). Ces journées ont permis de renforcer les contacts entre informaticiens et biologistes Lorrains.

2010-2014 : Participation au projet ANR PEPSI (Appel Modèles Numériques) coordonné par Sergei Grudinin (INRIA Grenoble Rhône-Alpes). Le projet PEPSI (Polynomial Expansions of Protein Structures and Interactions) regroupe trois partenaires : INRIA Grenoble Rhône-Alpes, équipe NANO-D, INRIA Nancy-Grand Est, équipe Orpailleur, Institut de Biologie Structurale - UMR Groupe Transporteurs membranaires. L'objectif du projet est de développer des moyens efficaces de représenter et de manipuler les structures tridimensionnelles (3D) de molécules de protéines et de calculer de manière fiable comment de grandes protéines interagissent. L'objectif principal est de développer de nouveaux outils et algorithmes qui seront utiles en médecine et en pharmacologie. Ma participation est liée à la thèse d'Anisah Ghoorah (encadrée par M-D. Devignes et Dave Ritchie) qui porte sur l'aide au *docking* protéine-protéine²¹ grâce à des modèles (templates) construits sur la base des interactions domaine-domaine connues et sur la classification des sites de liaisons par familles de domaines.

J'ai participé à quelques comités de programme de conférences nationales ou internationales :

- ACM SIGIR 2010 (33rd Annual ACM SIGIR Conference)
- DILS 2008, 5th international workshop on Data Integration in Life Sciences
- CORIA 2004 à 2014 (CONFérence en Recherche d'Information et ses Applications)
- Congrès Inforsid 1999, workshop *Data mining in Life Sciences* organisé lors des CONFérence ICDM (Industrial Conference on Data Mining) 2006 et 2007
- Workshop Case-Based Reasoning in the Health Sciences, organisé lors d'ICCBR 2007 (International Conference on Case-Based Reasoning)
- Workshop OGSB 2005 et 2006 (Ontologie, Grilles et intégration Sémantique pour la Biologie) organisé conjointement avec la conférence JOBIM 2005 et JOBIM 2006
- RED 2008, 1st International Workshop on REsource Discovery

Jurys de thèses : Claude Chellala, Adrien Coulet, Sid-Ahmed Benabderrahmane, Anisah Ghoorah, Emmanuel Bresso

Projet Européen Aquarelle (1996-1998) : Participation scientifique au projet Aquarelle. En 1995, l'INRIA Lorraine a participé à la définition du projet Aquarelle, qui a fait l'objet d'une réponse à un appel d'offre Ingénierie de l'Information du 4ème PCRD sous la responsabilité d'Alain Michard. Le projet Langue et Dialogue a assuré la coordination Nancéenne du

21. Le *docking* ou amarrage protéine-protéine est l'étude de la façon dont les protéines s'assemblent pour former des complexes capables de réaliser des fonctions biologiques.

projet et notamment la définition de trois work packages autour des thèmes suivants : (i) l'indexation automatique d'œuvres artistiques à partir de textes décrivant celles-ci ; (ii) l'utilisation d'alignements multilingues pour l'aide à la traduction de thésaurus spécialisés ; (iii) l'indexation semi-automatique d'images par l'utilisation de techniques de vision par ordinateur. J'ai pris en charge avec Antoine Tabbone du LORIA, Isabelle Gagliardi et Raymondo Schettini du CNR (Centre de Milan) le dernier workpackage. L'objectif était de concevoir et de prototyper un système multimédia distribué donnant accès à des documents relevant de l'héritage culturel européen. Notre contribution a porté sur la recherche d'images par le contenu et nous avons fourni un prototype prouvant la faisabilité d'une indexation semi-automatique d'images en utilisant des techniques de traitement d'images. Grâce à ce projet, nous avons eu l'opportunité d'acquérir une base de 2500 images pour valider le prototype développé par Gérard Duffing durant sa thèse soutenue en 2004.

5 Responsabilités administratives : participation aux conseils et mandats nationaux

2006-2010 Membre élue du conseil scientifique de l'UFR STMIA (Sciences & Techniques, Mathématiques Informatique et Automatiques) puis au conseil scientifique du secteur MIAE (Mathématiques Informatique Automatique Électronique) de la faculté des Sciences et Technologies de l'Université de Lorraine (Trois UFR ont été groupées en une en 2010).

2007-2011 Membre du CNU (Conseil National des Universités), section 27, élue sur la liste SPECIF.

2008-2011 Membre de la CDT (Commission de Développements Technologiques) d'Inria Nancy Grand-Est. Cette commission est chargée d'évaluer les demandes d'actions de développement technologiques à des fins de transfert ou de support à la recherche. Les avis rendus par la CDT servent à éclairer et à faciliter les arbitrages de la direction du centre Inria quant à l'attribution et le renouvellement de contrats d'ingénieurs jeunes diplômés et plus généralement toute demande de soutien pour des actions de développement technologique (ADT).

6 Résumé des activités pédagogiques

Je suis responsable de 5 modules d'enseignements du niveau L2 au niveau M1. Mes enseignements portent principalement sur la conception de systèmes d'information (niveaux L3 et 2ème année d'école d'ingénieurs), les langages associés au modèle relationnel, les extensions procédurales de SQL et son immersion dans des langages de programmation tels que Java ou C (niveau L2 informatique). J'avais enseigné, lors des six premières années de mon activité, les principes de la recherche documentaire et de la recherche d'information (modèles d'indexation/d'appariement/évaluation, bouclage de pertinence, moteurs de recherche...) à divers publics : Maîtrise du cinéma et de l'audio-visuel (Institut Européen du Cinéma de Nancy), Licence en sciences du langage (université Nancy 2), DESS Imagerie numérique et Interactivité. J'ai assuré pendant 4 ans les TD/TP d'un module sur les langages autour du méta-langage de balisage XML (DTD puis schémas, Xpath, XSL, Xquery) en 2ème année d'école d'ingénieurs (ESIAL). J'ai assuré les TD/TP d'un module consacré aux aspects avancés des SGBD (concurrence, confidentialité, optimisation de requêtes, gestion des transactions).

J'ai régulièrement animé, depuis 1998, des équipes pédagogiques composées de 6 à 8 personnes intervenant dans les enseignements dont je suis responsable incluant des collègues débutants ou non, des ATER, des moniteurs et des vacataires. En plus des enseignements en présentiel, j'ai encadré de nombreux projets tutorés, de stages de DEA ou de Master 2, deux mémoires CNAM d'une durée de 1 an et quelques projets d'initiation à la recherche d'élèves ingénieurs (ESIAL devenue Télécom Nancy).

Du fait de l'orientation de mon activité de recherche vers la bio-informatique, j'ai eu l'occasion d'enseigner à des étudiants biologistes (maîtrise de Génétique Cellulaire et Moléculaire et DESS RGTI puis Master 2). J'ai ainsi monté des cours d'initiation aux bases de données (conception et développement) pour débutants en privilégiant une mise en pratique avec des données et des exemples biologiques. J'ai également monté un cours sur les BD biologiques publiques avec des grilles de classification originales et la proposition d'une démarche pour l'utilisation rationnelle de ces ressources. J'ai pu constater, avec le temps, que nos collègues biologistes ignoraient souvent l'existence de banques de données ou de programmes qui peuvent pourtant leur être très utiles. J'ai réfléchi, avec ma collègue M-D. Devignes, à des enseignements en formation continue à destination de nos collègues en sciences de la vie.

J'ai également, dans le cadre de Spécialité *Interaction, Perception, Apprentissage et Connaissance* du Master 2 Informatique portée par Bernard Girau, proposé deux unités d'enseignement portant sur la fouille de données biomédicales et sur la mise en œuvre concrète du processus d'ECD. A la rentrée 2014, je serai responsable d'un module intitulé *Fouille de données et extraction de connaissances* dans la spécialité *Ingénierie et Applications des Masses de Données* de l'école d'ingénieurs TELECOM Nancy. Vu les enjeux économiques et stratégiques de l'exploitation des gros volumes de données, je suis convaincue que ce type d'enseignement sera petit à petit intégré dans l'ensemble des formations relevant de l'informatique.

En termes de responsabilités de formation, j'ai été co-responsable pendant 3 ans (2003-2006) du DESS RGTI *Ressources Génomiques et Traitements Informatiques* puis de la spécialité *Génomique et Informatique* du Master 2 Professionnel *Sciences de la Vie et de la Santé*. Cette formation s'adressait principalement à des biologistes spécialisés en génétique mais aussi à des informaticiens souhaitant acquérir une double-compétence dans le domaine de la génomique. Le responsable en était Pierre Leblond, Professeur à l'UHP en Biologie Cellulaire. J'ai assuré la coordination et l'évolution des enseignements informatiques et bioinformatiques et j'ai géré l'emploi du temps de la formation. Les effectifs des promotions ont fluctué autour d'une douzaine d'étudiants. Pierre Leblond et moi-même avons également procédé à l'examen des dossiers de candidature et aux auditions des candidats. Nous avons fait partie de tous les jurys de soutenances de projets tutorés et de stages de fin d'études.

Je suis actuellement responsable de l'enseignement de la bioinformatique et de trois unités d'enseignement dans le master CMI (Cursus Master Ingénierie) Biologie, Santé et Environnement (BSE).

7 Principales publications au 1er avril 2014

Articles dans des revues internationales avec comité de lecture

1. Caradec, Thibault ; Pupin, Maude ; Vanvlassenbroeck, Aurélien ; Devignes, Marie-Dominique ; Smail-Tabbone, Malika ; Jacques, Philippe, Leclère, Valérie, *Prediction of monomer isomery in florine : A workflow dedicated to nonribosomal peptide discovery*, PLoS ONE, 9(1) :e85667, 01 2014.

2. Bresso, Emmanuel ; Grisoni, Renaud ; Marchetti, Gino ; Karaboga, Arnaud Sinan ; Souchet, Michel ; Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika, *Integrative relational machine-learning approach for understanding drug side-effect profiles*, BMC Bioinformatics, 14 :207 (2013)
3. Devignes, Marie-Dominique ; Benabderrahmane, Sidahmed ; Smaïl-Tabbone, Malika ; Napoli, Amedeo ; Poch, Olivier, *Functional classification of genes using semantic distance and fuzzy clustering approach : evaluation with reference sets and overlap analysis*, International Journal of Computational Biology and Drug Design. Special Issue on : *Systems Biology Approaches in Biological and Biomedical Research*, 5 :3/4, p.245-260 (2012)
4. Ghoorah, Anisah ; Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika ; Ritchie, David, *Spatial clustering of protein binding sites for template-based protein docking*, Bioinformatics 27/20, p.2820-2827 (2011)
5. Benabderrahmane, Sidahmed ; Smaïl-Tabbone, Malika ; Poch, Olivier ; Napoli, Amedeo ; Devignes, Marie-Dominique, *IntelliGO : a new vector-based semantic similarity measure including annotation origin*, BMC Bioinformatics, 11 :588 (2010)
6. Ghemtio, Leo ; Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika ; Souchet, Michel ; Leroux, Vincent ; Maigret, Bernard, *Comparison of three preprocessing filters efficiency in virtual screening : identification of new putative LXRbeta regulators as a test case*, Journal of chemical information and modeling, 50/5, p.701-715 (2010)
7. Devignes, Marie-Dominique ; Franiatte, Philippe ; Messai, Nizar ; Bresso, Emmanuel ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *BioRegistry : Automatic extraction of metadata for biological database retrieval and discovery*, International Journal of Metadata Semantics and Ontologies, 5/3, p.184-193 (2010)
8. Yilmaz, Saliha ; Jonveaux, Philippe ; Bicep, Cedric ; Pierron, Laurent ; Smaïl-Tabbone, Malika ; Devignes, Marie-Dominique, *Gene-Disease Relationship Discovery based on Model-driven Data Integration and Database View Definition*, Bioinformatics, 25/2, p.230-236 (2009)
9. Beautrait, Alexandre ; Leroux, Vincent ; Chavent, Matthieu ; Ghemtio, Léo ; Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika ; Cai, Wensheng ; Shao, Xuegang ; Moreau, Gilles ; Bladon, Peter ; Yao, Jianhua ; Maigret, Bernard, *Multiple-step virtual screening using VSM-G : overview and validation of fast geometrical matching enrichment.*, Journal of Molecular Modeling, 14/2, p.135-148 (2008)
10. Coulet, Adrien ; Smaïl-Tabbone, Malika ; Benlian, Pascale ; Napoli, Amedeo ; Devignes, Marie-Dominique, *Ontology-guided data preparation for discovering genotype-phenotype relationships*, BMC Bioinformatics, 9 Suppl 4 S3 (2008)

Articles dans des revues nationales avec comité de lecture

1. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *Correction et complétude d'un algorithme de recherche d'information par treillis de concepts*, Revue des Nouvelles Technologies de l'Information RNTI (2007)
2. Devignes, Marie-Dominique ; Schaaff, André ; Smaïl, Malika, *Collecte et intégration de données biologiques hétérogènes sur le web : application dans le domaine de la cartographie du génome humain*, Ingénierie des Systèmes d'Information, 7/1-2, p.45-61 (2002)
3. Smaïl, Malika, *Vers des systèmes évolutifs de recherche d'information : un état de l'art*, Technique et Science Informatiques - TSI, 17/10, p.1193-1222 (1998)

Chapitres d'ouvrages scientifiques

1. Coulet, Adrien ; Smaïl-Tabbone, Malika ; Napoli, Amedeo ; Devignes, Marie-Dominique, *Ontology-based knowledge discovery in pharmacogenomics*, Advances in Experimental Medicine and Biology, Volume 696, Part 5, Springer, p.357-366 (2011)
2. Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika, *Maîtriser les ressources numériques : biologie in silico*, Biologie L'ère numérique, CNRS Editions, Magali Roux (Eds), p.189-222 (2009)
3. Simonnot, Brigitte ; Smaïl, Malika, *Model for interactive retrieval of videos and still images*, Bruce Berra, Kingsley C. Nwosu and Bhavani Thuraisingham (Eds), Kluwer Publishers, p.278-317 (1996)

Conférences internationales avec actes et comité de lecture

1. Personeni, Gabin ; Daget, Simon ; Bonnet, Céline ; Jonveaux, Philippe ; Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika ; Coulet, Adrien, *Mining Linked Open Data : a Case Study with Genes Responsible for intellectual disability*, In Proc. of the 10th International Conference on Data Integration in the Life Sciences 2014 - DILS 2014, volume to appear of Lecture Notes in Bioinformatics, Springer, 15 pages (2014)
2. Bresso E., Grisoni R., Devignes M.-D., Napoli A., and Smaïl-Tabbone M., *ILP Characterization of 3D Protein-Binding Sites and FCA-Based Interpretation*, A. Fred, K. Dietz, J.L.G. and Liu, and J. Filipe, (editors), 4th International Joint Conference, IC3K 2012, Barcelona, Spain, October 4-7, 2012. Revised Selected Papers, volume 415 of Communications in Computer and Information Science, p.84-100. Springer (2013)
3. Alam M., Coulet A., Napoli A., Smaïl-Tabbone M., *Formal Concept Analysis Applied to Transcriptomic Data*, Ninth International Conference on Concept Lattices and Their Applications - CLA 2012, Spain, 6 pages (2012).
4. Bresso E., Benabderrahmane S. ; Smaïl-Tabbone M., Marchetti G. Karaboga A., Souchet M., Napoli A., Devignes M.-D., *Use of domain knowledge for dimension reduction : application to mining of drug side effects*, In Proc. 3rd International Conference on Knowledge Discovery and Information Retrieval - KDIR 2011, France. INSTICC, SciTePress Digital Library, 8 pages (2011)
5. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *Using Domain Knowledge to Guide Lattice-based Complex Data Exploration*, In Proc. 19th European Conference on Artificial Intelligence - ECAI 2010, Lisbon, Portugal p.847-852 (2010)
6. Ghemtio, Léo ; Smaïl-Tabbone, Malika ; Devignes, Marie-Dominique ; Souchet, Michel ; Maigret, Bernard, *A KDD Approach for Designing Filtering Strategies to Improve Virtual Screening*, KDIR 2009 - International Conference on Knowledge Discovery and Information Retrieval, Madeira Portugal, INSTICC, SciTePress Digital Library, 8 pages (2009)
7. Devignes, Marie-Dominique ; Franiatte, Philippe ; Messai, Nizar ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *BioRegistry : Automatic Extraction of Metadata for Biological Database Retrieval and Discovery*, 1st international workshop on Ressource Discovery (RED), Joint to iiWAS 2008, Linz, Austria (2008)
8. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *Extending Attribute Dependencies for Lattice-Based Querying and Navigation*, Peter W. Ek-lund and Ollivier Haemmerlé (Eds), Lecture Notes in Computer Science 5113 (Springer),

- In Proc. 16th International Conference on Conceptual Structures - ICCS 2008, Toulouse, France p.189-202 (2008)
9. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *Many-Valued Concept Lattices for Conceptual Clustering and Information Retrieval*, Malik Ghalab (Eds) 18th European Conference in Artificial Intelligence - ECAI 2008, IOS Press, Patras, Greece p.127-131 (2008)
 10. Coulet, Adrien ; Smaïl-Tabbone, Malika ; Benlian, Pascale ; Napoli, Amedeo ; Devignes, Marie-Dominique, *Ontology-guided Data Preparation for Discovering Genotype-Phenotype Relationships*, Network Tools and Applications in Biology : A Semantic Web for Bioinformatics - NETTAB 2007 Pisa, Italy (2007)
 11. Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika, *Workshop PC Chairs' Message. Web Data Integration for Mining in the Life Sciences (WebDIM4LS)*, Mathias Weske and Mohand-Said Hacid and Claude Godart (Eds) Lecture Notes in Computer Science 4832 (Springer), International Workshops on Web Information Systems Engineering - WISE 2007 Workshops, Nancy, France p.3-4 (2007)
 12. Coulet, Adrien ; Smaïl-Tabbone, Malika ; Benlian, Pascale ; Napoli, Amedeo ; Devignes, Marie-Dominique, *SNP-Converter : an Ontology-Based solution to Reconcile Heterogeneous SNP Descriptions for Pharmacogenomic Studies*, 3rd International Workshop on Data Integration in the Life Sciences 2006 - DILS'06, European Bioinformatics Institute (EBI), Hinxton/UK (2006)
 13. Coulet, Adrien ; Smaïl-Tabbone, Malika ; Napoli, Amedeo ; Devignes, Marie-Dominique, *Suggested Ontology for Pharmacogenomics (SO-Pharm) : Modular Construction and Preliminary Testing*, Proceedings of International Workshop on Knowledge Systems in Bioinformatics - KSinBIT'06, Montpellier France (2006) (2006)
 14. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *BR-Explorer : An FCA-based algorithm for Information Retrieval*, Fourth International Conference On Concept Lattices and Their Applications - CLA 2006, Hammamet/Tunisia (2006)
 15. Smaïl-Tabbone, Malika ; Osman, Shazia ; Messai, Nizar ; Napoli, Amedeo ; Devignes, Marie-Dominique, *BioRegistry : a structured metadata repository for bioinformatic databases*, M.R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fisher (Eds) Lecture Notes in Bioinformatics 3695 (Springer), First International Symposium on Computational Life Science - CompLife 2005, Konstanz, Germany, p.46-56 (2005)
 16. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *Querying a Bioinformatic Data Sources Registry with Concept Lattices*, Frithjof Dau, Marie-Laure Mugnier and Gerd Stumme (Eds) LNAI 3596 (Springer), 3rd International Conference on Conceptual Structures - ICCS 2005, Kassel, Germany, p.323-336 (2005)
 17. Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika, *Integration of biological data from web resources : management of multiple answers through metadata retrieval*, 12th International Conference on Intelligent Systems for Molecular Biology - 3rd European Conference on Computational Biology - ISMB-ECCB 2004, Glasgow, Scotland, 3 pages (2004)
 18. Duffing, Gérald ; Smaïl, Malika, *Organising and Searching Partially Indexed Image Databases*, 24th BCS-IRSG European Colloquium on Information Retrieval Research, Springer Verlag, Glasgow, Scotland, 19 pages (2002)
 19. Duffing, Gérald ; Smaïl, Malika, *A Novel Approach for Accessing Partially Indexed Image Corpora*, 4th International Conference on Visual Information Systems - VISUAL'2000, Lyon, France, 13 pages (2000)

Conférences nationales avec actes et comité de lecture

1. Hafiane, Rachid ; Smaïl-Tabbone, Malika ; Devignes, Marie-Dominique ; Tabbone, Salvatore, *Clustering optimal de gènes fondé sur une mesure de similarité sémantique*, Actes de la Conférence en Recherche d'Information et Applications - CORIA 2013, Neuchâtel, Suisse, 15 pages (2013)
2. Pupin, Maude ; Smaïl-Tabbone, Malika ; Jacques Philippe ; Devignes Marie-Dominique ; Leclère, Valérie, *NRPS toolbox for the discovery of new nonribosomal peptides and synthetases*, Actes des Journées Ouvertes en Biologie, l'Informatique et les Mathématiques - JOBIM 2012, Rennes, France, p.123-127 (2012)
3. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *Connaissances de domaine et treillis de concepts pour l'exploration progressive de données complexes*, Actes des 21ème Journées francophones d'Ingénierie des Connaissances, IC 2010, Nîmes, France, p.233-244 (2010)
4. Benabderrahmane, Sidahmed ; Devignes, Marie-Dominique ; Smaïl-Tabbone, Malika ; Poch, O. ; Napoli, Amedeo ; Nguyen N.-H, N. ; Raffelsberger, W., *Analyse de données transcriptomiques : Modélisation floue de profils d'expression différentielle et analyse fonctionnelle.*, Actes du congrès Informatique des Organisations et Systèmes d'information et de décision - INFORSID 2009 Toulouse, France, p.413-428 (2009)
5. Messai, Nizar ; Devignes, Marie-Dominique ; Napoli, Amedeo ; Smaïl-Tabbone, Malika, *Traitement d'attributs inter-dépendants pour la recherche d'information par treillis*, Cépaduès éditions, Actes des journées francophones d'Ingénierie des Connaissances - IC 2007 Grenoble, France, p.109-120 (2007)
6. Smaïl, Malika, *Recherche de régularités dans une mémoire de sessions de recherche d'information documentaire*, Actes du congrès Informatique des Organisations et Systèmes d'information et de décision - INFORSID 1999, La Garde, Toulon, 19 pages (1999)