Projet ANR (2016-2021) “ PractiKPharma ”: extraction, comparaison et découverte de connaissances en pharmacogénomique

Pierre Monnin, Adrien Coulet

Projet ANR (2016-2021) « PractiKPharma » : extraction, comparaison et découverte de connaissances en pharmacogénomique

Pierre Monnin [1] et Adrien Coulet [2]

Introduction

Le projet ANR PractiKPharma [3] (2016 – 2021) s'est intéressé au développement d'approches informatiques pour le domaine de la pharmacogénomique (PGx). Ce domaine étudie l'influence des variations génétiques des individus sur leur réponse aux médicaments. En d'autres termes, la PGx s'intéresse aux relations illustrées par la figure 1 qui lient un ensemble de médicaments, un ensemble de facteurs génétiques, et un ensemble de réponses aux médicaments (effets attendus, indésirables, ou absence d'effet). Par exemple, les relations PGx représentées en figure 2 indiquent qu'un individu traité avec de la codéine pourra connaître une absence d'effet, l'effet analgésique attendu, ou un effet indésirable de toxicité en fonction du variant génétique porté par le gène CYP2D6.

L'état de l'art en PGx est conséquent mais très inégalement validé car une grande partie des observations sont peu ou pas reproduites. Par exemple, la figure 3 illustre que 90 % des relations PGx représentées dans PharmGKB [23], base de données

1. Chercheur, Orange, Belfort, pierre.monnin@orange.com.

2. Chercheur, équipe HeKA, Inria Paris, centre de recherche des Cordeliers, Inserm, université de Paris, adrien.coulet@inria.fr.

3. http://practikpharma.loria.fr

DOI:10.48556/SIF.1024.19.109

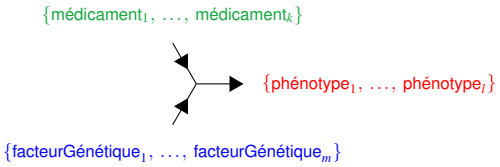

FIGURE 1. Modèle abstrait d'une relation pharmacogénomique.

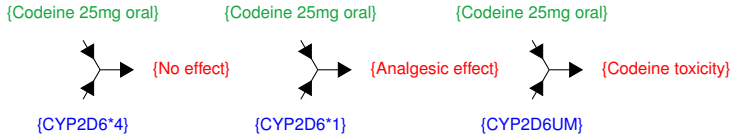

FIGURE 2. Relations pharmacogénomiques décrivant l'influence de trois variants du gène CYP2D6 sur la réponse à un même traitement de codéine. Les patients ayant le variant CYP2D6*4 ne connaîtront pas l'effet analgésique attendu du traitement ; ceux ayant le variant CYP2D6UP connaîtront une toxicité du traitement.

de référence du domaine, ne sont pas directement applicables en pratique clinique par manque de validation. Le but de PractiKPharma a été de fournir des méthodes et outils de gestion de connaissances pour progresser dans la validation de ce qui ne l'est pas, en extrayant et comparant des connaissances issues de sources diverses comme les bases de données spécialisées, la littérature biomédicale et les dossiers patients informatisés (DPI). Dans cet objectif, le projet a suivi quatre axes de travail illustrés en figure 4 et décrits ci-dessous :

(1) l'extraction de connaissances de l'état de l'art à partir de bases de données spécialisées et de la littérature ;

(2) l'extraction de connaissances « observationnelles » à partir des DPI pour identifier les connaissances pouvant être mises en œuvre en médecine personnalisée ;

(3) la comparaison des connaissances extraites en (1) et (2) ;

(4) la valorisation des connaissances extraites en cherchant des relations gènes–médicaments plus probables et des mécanismes moléculaires capables d'expliquer la survenue d'effets indésirables. Ce dernier objectif est motivé par le fait que les mécanismes mis en jeu dans les réponses aux médicaments sont très souvent inconnus. Or, la connaissance des facteurs génétiques associés aux réponses peut servir de point d'entrée à l'identification des mécanismes moléculaires sous-jacents.

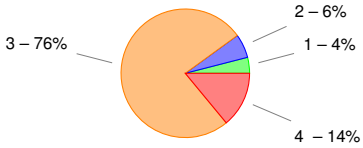

FIGURE 3. Répartition des niveaux de validation des connaissances pharmacogénomiques dans la base de référence PharmGKB (au 05/07/2019). Les niveaux 1 et 2 correspondent à des connaissances implémentées en pratique clinique ou supportées par des études montrant un niveau fort (1) ou modéré (2) d'association. Les niveaux 3 et 4 correspondent quant-à-eux à des connaissances décrites dans des études non-répliquées, dans de multiples études montrant un manque de preuve, ou dans des études non-significatives. 90% (3+4) des connaissances nécessitent plus de validation pour pouvoir être utilisées en clinique.

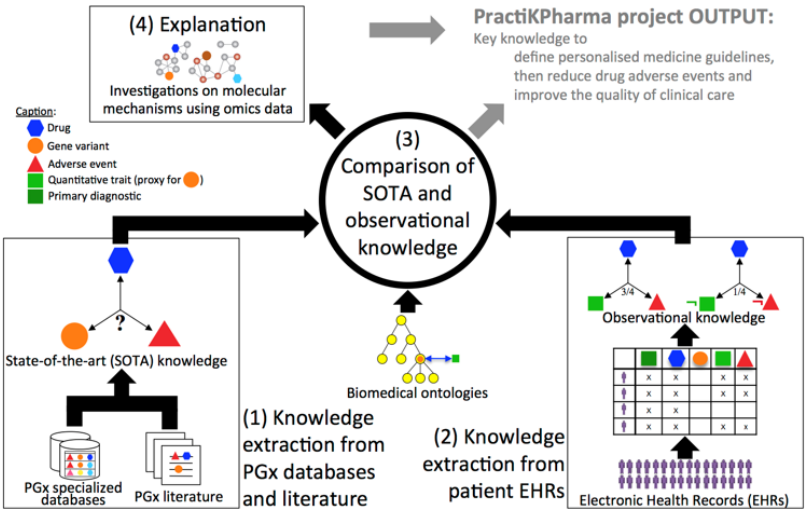

FIGURE 4. Les quatre axes de travail du projet PractiKPharma

PractiKPharma a été mené par :

— des experts en pharmacie et pharmacovigilance du service SSPIM du CHU de Saint-Étienne et du centre régional de pharmacovigilance du CHRU de Nancy ;

— des experts en gestion de dossiers patients informatisés de l'hôpital Georges Pompidou de l'AP-HP ;

— des experts en gestion de connaissances et d'ontologies du laboratoire LIRMM de Montpellier ;

— des experts en gestion et extraction de connaissances du Loria de Nancy.

Principaux résultats

Extraction de connaissances pharmacogénomiques à partir de l'état de l'art

L'objectif de cet axe de travail est de permettre une extraction automatique des connaissances de l'état de l'art en pharmacogénomique (PGx) et de les intégrer dans une plateforme qui permet leur comparaison et leur analyse. Nous avons extrait des connaissances de deux types de sources : des textes, avec les résumés d'articles scientifiques de la base de données PubMed, et des données structurées et parfois semi-structurées de la base de données PharmGKB [23], référence en PGx. Les connaissances extraites (et les entités impliquées comme les gènes et les phénotypes) ont d'abord été normalisées à l'aide de vocabulaires spécialisés du domaine (ici appelés ontologies) avant d'être intégrées dans PGxLOD [4] [16], un graphe de connaissances que nous avons construit en suivant les principes du Web Sémantique [3] et les principes FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) [24].

Il est important de noter qu'au commencement du projet PractiKPharma, il n'existait pas de corpus annoté dédié au domaine de la PGx. En particulier, il n'existait pas de corpus avec des phrases simultanément annotées par les trois types d'entités d'intérêt en PGx : facteur génétique (gène, variant, haplotype, etc.), médicament, et phénotype de réponse au médicament. L'absence d'un tel corpus a guidé nos efforts dans deux directions suivies de front :

(1) l'apprentissage par transfert pour l'extraction de relations PGx, avec l'utilisation de corpus existants mais connexes à notre tâche (où, par exemple, seuls un ou deux des types d'entités d'intérêts sont annotés) ;

(2) la création d'un corpus manuellement annoté, dédié à notre tâche et donc incluant les trois types d'entités.

Considérant l'apprentissage par transfert (1) pour la tâche d'extraction de relations à partir de textes, nous avons montré qu'enrichir un corpus cible (où les relations annotées sont celles que l'on cherche à extraire) de petite taille avec un corpus source de plus grande taille (où les relations annotées sont d'un type distinct) améliore les performances, en particulier lorsque le modèle d'extraction de relations

4. https://pgxlod.loria.fr

peut considérer des informations sur la syntaxe des phrases [14]. Nous avons en particulier exploré l'hypothèse selon laquelle les modèles pourraient gagner en performances en généralisant des connaissances sur la syntaxe de l'expression des relations en anglais, même si celles-ci sont de types différents. Cette hypothèse semble valide dans notre contexte applicatif et mériterait davantage d'expérimentations pour être généralisée.

Considérant la constitution d'un corpus inédit (2), nous avons assemblé et ouvert PGxCorpus [5] [13], un corpus de 945 phrases issues de résumés d'articles scientifiques où les entités d'intérêt en PGx et leurs relations ont été annotées manuellement par 11 annotateurs. Le corpus final contient 2875 relations annotées, chacune ayant été vue par au moins quatre annotateurs différents. Nous avons dans un premier temps montré l'utilité de ce nouveau corpus pour l'extraction automatique de relations [13], puis nous avons montré qu'un modèle de type BERT, pré-entraîné avec des textes biomédicaux multi-domaines et réglé finement avec PGxCorpus, permet de dépasser les meilleures performances de l'état de l'art pour la tâche cible [12]. Nous estimons que PGxCorpus permettra d'une part de progresser dans la gestion des connaissances associées à ce domaine, notamment en permettant d'entraîner ou de régler finement des modèles supervisés. D'autre part, PGxCorpus contient plusieurs particularités linguistiques constituant des challenges en traitement automatique de la langue (TAL) comme des entités discontinues, imbriquées, et des relations ternaires. Nous pensons que PGxCorpus permettra d'évaluer des outils de TAL génériques pour l'extraction de ces entités ou relations particulières.

Concernant l'extraction de connaissances à partir de bases de données expertes, nous avons développé des scripts d'extraction automatique des variants, médicaments, et réponses aux médicaments décrits dans PharmGKB, ainsi que de leurs relations disponibles de façon structurée ou semi-structurée dans la base. Cette extraction ne constitue pas une contribution scientifique en tant que telle, mais le résultat de l'extraction vient compléter notre graphe de connaissances appelé PGxLOD et constitue la version la plus récente de PharmGKB en RDF disponible à la communauté depuis que le projet Bio2RDF n'est plus maintenu [11].

L'ensemble des relations extraites de la littérature ainsi que les relations extraites de PharmGKB sont structurées selon une ontologie minimale appelée PGxO et regroupées au sein d'un graphe de connaissances appelé PGxLOD [16]. En plus des relations PGx, PGxLOD regroupe des connaissances associées aux gènes, médicaments et phénotypes en intégrant notamment le contenu des bases de données ClinVar, DrugBank, SIDER, et CTD. PGxLOD est une ressource qui suit les préceptes du Web Sémantique et respecte les principes FAIR. Notons, par exemple, que chaque connaissance représentée dans ce graphe est associée à une provenance bien définie. PGxLOD est indexé dans Google Dataset Search et LOD Cloud. Ce graphe de

5. https://pgxcorpus.loria.fr.

connaissances constitue la plateforme expérimentale du projet pour la comparaison de connaissances PGx de provenances différentes (Section) et la recherche d'éléments explicatifs et mécanistiques quant à la survenue d'effets indésirables (Section).

Extraction de connaissances pharmacogénomiques à partir des dossiers patients

L'objectif initial de cet axe de travail est d'extraire des connaissances pharmacogénomiques (PGx) à partir de dossiers patients informatisés (DPI). Parmi les tâches prévues, nous souhaitions constituer une cohorte de patients et mesurer la variabilité de leur réponse à un médicament particulier. L'absence de données génétiques dans les DPI était un obstacle identifié dès le départ que nous pensions surmonter en mettant en évidence des *surrogates*, c'est à dire des co-variables mesurées de façon routinières dans les DPI et qui puissent être associées statistiquement à des variants génétiques (non mesurés en routine).

L'accessibilité, le contenu et la qualité des données des DPI ont orienté nos contributions vers une extraction de connaissances à partir des textes cliniques associées aux DPI. Pour cela, nous avons avancé le développement d'outils pour la reconnaissance d'entités nommées dans les textes cliniques et pour la détection de leur contexte : sont-elles niées ? Concernent-elles le patient ou un membre de sa famille ? Le présent ou le passé ? Sont-elles dans une zone dupliquée de compte rendu en compte rendu ? French Annotator, NCBO Annotator+, FrenchFast Context et d'autres approches développées dans le cadre du projet permettent de mieux répondre à ces questions en français [9, 15, 21, 22]. Ces éléments contextuels sont cruciaux afin de limiter le nombre de faux positifs que les outils de reconnaissances d'entités nommées produisent s'ils sont utilisés seuls. En parallèle, nous avons évalué et comparé la facilité de réutilisation des suites logiciels standards pour le traitement automatique du langage (TAL) dans le cas particulier des textes cliniques hospitaliers [10] et nous avons participé au développement d'une librairie appelée PyMedExt dont l'objectif est de faciliter le traitement des textes cliniques [6]. Nous avons développé une plateforme à l'aide de Galaxie, un gestionnaire de *workflows* bioinformatiques, qui facilite la gestion de données génomiques des patients par les biologistes moléculaires de l'HEGP [8]. Nous voyons ces différentes contributions comme des briques logicielles nécessaires pour permettre dans le futur l'extraction de connaissances PGx à partir de DPI.

Dans le cadre d'une collaboration avec Stanford, dont les notes cliniques en anglais sont déjà annotées en tenant compte du contexte, nous avons pu mener deux travaux d'analyse autour de la variabilité individuelle de la réponse aux médicaments. Nous avons utilisé des extensions de l'analyse formelle de concepts (AFC) pour trouver des ensembles de réactions indésirables aux médicaments souvent observées chez les mêmes patients [20]. Et, poussé par l'absence de données génétiques, nous

6. Résultat non encore publié, `https://github.com/equipe22/pymedext_core`.

avons expérimenté l'utilisation des changements de doses de traitement comme marqueur des profils de réponse, en faisant l'hypothèse qu'une réduction de dose est le signe d'une surréaction à un traitement. Dans ce contexte, nous avons montré pour une vingtaine de médicaments connus en PGx que nous pouvions, avec des données récoltées de façon routinière dans les DPI, prédire si un patient bénéficierait d'une réduction de dose pour éviter un effet indésirable et cela avant que le médicament ne soit prescrit [5].

Les résultats obtenus sont conséquents et illustrent l'intérêt d'utiliser les DPI. Néanmoins, ils demeurent en décalage avec l'objectif premier, trop ambitieux, d'extraction de connaissances PGx à partir de ces données complexes. En réponse à ce constat, nos contributions ont visé d'une part à faire face au premier challenge que constitue l'utilisation des textes des DPI ; et d'autre part à prendre du recul, pour nous intéresser à la variabilité de réponse aux médicaments de façon plus générale, et pas seulement celle causée par la génétique.

Comparaison de connaissances

L'objectif de cet axe de travail est de développer des méthodes et outils pour permettre la comparaison des connaissances de provenances diverses. Nous avons expérimenté avec la pharmacogénomique (PGx) mais avons apporté une attention particulière à ce que ces méthodes soient les plus générales possibles.

Notre premier travail d'investigation pour la comparaison de connaissances a été de nous assurer que les vocabulaires et ontologies utilisés dans les différentes sources de connaissances considérées pouvaient être alignés. Nous avons observé que dans la plupart des cas, les alignements multilingues proposés dans [1] sont suffisants pour réconcilier nos données issues de l'état de l'art ("annotables" avec des ontologies anglo-saxonnes) et des DPI ("annotables" avec des ontologies francophones). Ces alignements multilingues mettent en évidence la relative pauvreté lexicale des ressources françaises, en comparaison des ressources anglo-saxonnes. Pour cette raison, nous avons réorienté les efforts de PractiKPharma sur le développement de méthodes d'alignement entre ontologies anglo-saxonnes. Dans cette optique, nous avons étudié l'usage de ressources externes (c'est-à-dire d'autres ontologies) pour l'alignement d'ontologies et proposé une méthode qui dépasse l'état de l'art dans ce domaine [2].

La très grande hétérogénéité des sources de données en termes de vocabulaire, granularité et niveau de discours, fait que la normalisation par ontologies des connaissances extraites précédemment n'est pas suffisante pour leur alignement. Par exemple, la figure 5 illustre l'hétérogénéité des connaissances extraites (différentes langues, différents vocabulaires, arguments inconnus) et des alignements attendus entre elles (connaissances identiques, plus spécifiques, similaires à un certain degré). Le second travail d'investigation s'est pour cette raison intéressé à la comparaison des connaissances regroupées au sein du graphe de connaissances PGxLOD, en tirant au maximum parti des connaissances de domaines associées (*i.e.*, des

ontologies). Dans ce contexte, nous nous sommes intéressés au cas particulier de l'alignement de relations *n*-aires dans un graphe de connaissances, de par l'exemple des relations PGx, qui sont définies par l'ensemble des individus qu'elles relient. L'alignement automatique de telles entités au sein d'un graphe de connaissances n'est pas quelque chose de classique et résolu dans l'état de l'art et nous y avons contribué par deux travaux. Notre premier travail propose une approche symbolique non supervisée, à base de règles formelles qui permettent d'identifier si deux individus du graphe qui ont des provenances distinctes sont équivalents, plus spécifiques, ou similaires, et cela au regard de leurs voisinages directs dans le graphe (*i.e.*, les composants associés par la relation *n*-aire) et des connaissances de domaines [17]. Notre second travail propose une approche supervisée numérique, qui apprend à partir des exemples de pairs d'individus du graphe à distinguer ceux qui sont équivalents, plus généraux, ou similaires [19]. Pour cette seconde approche nous avons appris une représentation de nos relations de façon supervisée avec un réseau convolutif de graphe (GCN pour *Graph Convolutional Network*). Suivant cette approche, nous avons particulièrement étudié la façon avec laquelle le GCN pouvait retrouver des types de similarités distincts à partir du voisinage. Ces deux méthodes, l'une symbolique et l'autre numérique, ont la qualité d'être relativement générales et applicables à d'autres domaines que la pharmacogénomique. Du point de vue du cas d'application de notre projet, nous les avons mis en œuvre pour aligner au sein de PGxLOD les connaissances de provenances diverses. Les nouveaux alignements obtenus sont intéressants, premièrement car ils mettent en lumière des relations qui sont décrites dans plusieurs publications mais sont absentes de PharmGKB (soulignant un manque dans la base) ; deuxièmement des relations présentes dans la base, en manque de références scientifiques pour documenter leur niveau de validation, sont mises en relation avec des publications qui mentionnent ces relations.

Dans cet axe de travail, nous avons proposé et mis en œuvre des méthodes originales de comparaison de connaissances et établi un système fonctionnel qui met en évidence les connaissances équivalentes ou similaires proposées dans différentes sources. Ceci est unique et ouvre des perspectives d'analyse de données plus globales et uniques pour le domaine. Parmi les limites qu'il est important de citer, nous n'avons pas comparé l'état de l'art et des résultats d'analyse de cohortes de DPI, ce qui est une tâche complexe. Elle nécessiterait la mise en place d'une source experte qui associe variant génétique et traits phénotypiques observés dans les DPI. Une telle source n'existe pas et les éléments de connaissances qui pourraient la peupler résultent d'analyses d'envergure appelées *phenome-wide association sudies*.

Explication d'effets secondaires aux médicaments

L'objectif de cet axe de travail est la fouille de graphes de connaissances biomédicales pour découvrir des éléments explicatifs de la survenue d'effets médicamenteux indésirables. En particulier, nous pensons que les graphes de connaissances comme

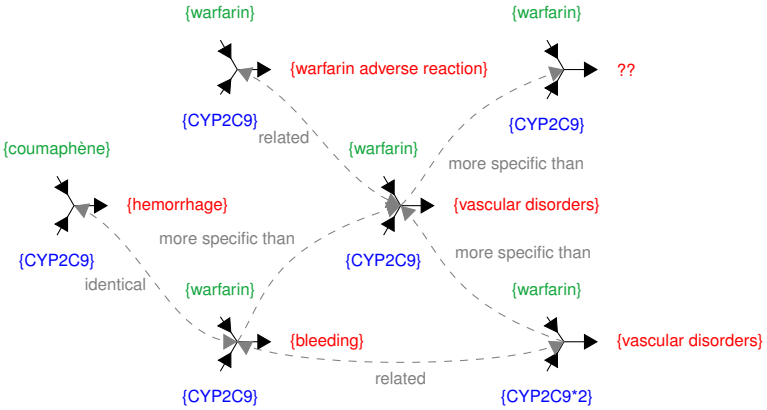

FIGURE 5. Exemples de relations pharmacogénomiques de provenances diverses et des alignements attendus entre elles. Leur hétérogénéité provient des faits suivants : un phénotype est inconnu (??) ; coumaphène est le terme français pour warfarin ; hemorrhage est un synonyme de bleeding ; CYP2C9*2 est un variant génétique plus précis que le gène CYP2C9 lui-même ; bleeding est plus spécifique que vascular disorder ; et vascular disorders est lié à warfarin adverse reaction.

PGxLOD associant des facteurs génétiques aux réponses aux médicaments et incluant de nombreux descripteurs des entités en jeu sont des sources de connaissances sous-exploitées pour la compréhension du mécanisme d'action des médicaments.

Dans un premier temps, PGxLOD a été enrichi avec des sources de données « omiques », telles que KEGG et CTD [16]. La taille de la version résultante de PGxLOD est de 88 millions de triplets [7]. L'exploration d'un graphe d'une telle taille soulève des problèmes de passage à l'échelle, en particulier, lorsque l'on souhaite considérer la sémantique associée au graphe de connaissances. Par exemple, des relations `partOf`, `owl:sameAs`, et `rdfs:subClassOf` sont transitives, et considérer cette propriété lors de la fouille du graphe complexifie la tâche. Pour cette raison, nous avons proposé une approche qui contrôle la quantité de chemins et de patrons de chemin à considérer lorsque l'on explore un graphe de connaissances dans une tâche de fouille [18]. Nous proposons de contrôler la quantité de chemin et patrons de chemin à la fois par des contraintes ad-hoc définies par l'analyste, mais également par des propriétés de monotonicité définies par les données. Nous avons utilisé cette

7. Les triplets ont la forme $\langle$sujet, prédicat, objet$\rangle$ et constituent les éléments de base des graphes de connaissances du Web Sémantique.

approche de fouille pour extraire de PGxLOD des chemins et patrons de chemins associés aux médicaments, et avons montré que ces chemins et patrons de chemins (1) sont de bons prédicteurs pour distinguer les médicaments qui causent un type d'effet indésirable des autres médicaments ; (2) constituent des éléments d'interprétation de la mécanique moléculaire sous-jacente à la survenue d'effets indésirables [4]. Pour cela, nous avons mis en œuvre des méthodes d'apprentissage supervisé symboliques simples mais nativement explicables que sont les arbres de décision et les règles de décision. Parmi les chemins et patrons de chemins qui sont des bons prédicteurs pour la classe de médicaments causant l'effet indésirable, nous sélectionnons ceux qui contiennent une entité de notre graphe potentiellement interprétable par une expert en pharmacologie (par exemple un terme *Gene Ontology*, le nom d'un réseau métabolique, ou celui d'un gène). Les prédicteurs sélectionnés ont ensuite été soumis à des experts en pharmacologie pour évaluer leur potentiel « explicatif », c'est-à-dire leur potentiel pour constituer une explication à la survenue de l'effet indésirable.

Conclusion et perspectives

Le projet PractiKPharma nous a permis de faire progresser l'état de l'art dans le domaine de la gestion des connaissances en pharmacogénomique et de produire des logiciels et ressources offrant des perspectives de nouveaux travaux et collaborations. Nous distinguons trois grands groupes de contributions : des contributions informatiques, d'ordre méthodologique, notamment autour de la comparaison de connaissances de provenances diverses ; des contributions applicatives sur l'extraction de connaissances à partir de textes cliniques avec le développement et partage d'outils fonctionnels ; des ressources ouvertes de référence pour la gestion des connaissances dans le domaine de la pharmacogénomique : un corpus de textes annotés (PGxCorpus) qui permet le développement d'outils d'extraction de connaissances plus performants que l'existant ; et un graphe de connaissances (PGxLOD) qui permet de comparer les connaissances du domaine.

Les ressources constituées au cours du projet ouvrent des perspectives d'un point de vue applicatif, mais également d'un point de vue informatique. En effet, PGxCorpus et PGxLOD offrent des terrains d'expérimentation respectivement pour le traitement naturel de langage (notamment autour des tâches de reconnaissance d'entités imbriquées ou discontinues et d'extraction de relations *n*-aires) et pour la comparaison de connaissances (notamment autour de l'alignement de relations *n*-aires, de l'inférence de liens). D'un point de vue applicatif, ces ressources offrent la possibilité de comparer les connaissances de l'état de l'art et d'identifier les éléments qui nécessitent plus de validation ou, au contraire, d'identifier des faisceaux concordants dans différente sources à propos d'un élément de connaissances, participant ainsi à sa confirmation. En perspective, citons notre volonté de connecter les connaissances de l'état de l'art représentées dans PGxLOD avec des données d'ordre observationnel

concernant des patients suivis à l'hôpital. Ainsi, un ou plusieurs patients ayant vécu une réponse médicamenteuse indésirable viendraient alors instancier une connaissances de l'état de l'art, permettant de valider (ou de modérer) les connaissances de l'état de l'art.

Le récent article de commentaires de Joshua Denny (*Chief Executive Officer of the National Institutes of Health's All of Us Research Program*) du 18 mars 2021 positionne la pharmacogénomique, l'utilisation des DPI, et l'intelligence artificielle comme des éléments clés dans la réalisation de la médecine de précision dans les années à venir [7]. L'expertise acquise dans ces domaines par les partenaires du projet nous positionne de façon favorable pour participer activement à cette réalisation via, notamment, notre participation à la création de l'équipe-projet HeKA (Inria, Inserm, université Paris) [6].

Remerciements

Ces travaux ont été soutenus par l'Agence nationale de la recherche dans le cadre du projet PractiKPharma (ANR-15-CE23-0028), par Inria dans le cadre de l'équipe-associée Inria-Stanford *Snowball*, et par l'IDEX « Lorraine université d'excellence » (15-IDEX-0004).

Références

[1] Amina Annane et al. Réconciliation d'alignements multilingues dans BioPortal. In *IC : Ingénierie des Connaissances*, June 2016.

[2] Amina Annane et al. Building an effective and efficient background knowledge resource to enhance ontology matching. *Journal of Web Semantics*, 51 :51–68, 2018.

[3] Tim Berners-Lee et al. The semantic web. *Scientific american*, 284(5) :28–37, 2001.

[4] Emmanuel Bresso et al. Investigating ADR mechanisms with explainable AI : a feasibility study with knowledge graph mining. *BMC Medical Informatics Decision Making*, 21(1) :171, 2021.

[5] Adrien Coulet et al. Predicting the need for a reduced drug dose, at first prescription. *Scientific Reports*, 8(1), October 2018.

[6] Adrien Coulet et al. L'équipe-projet HeKA. *Bulletin de l'Association Française pour l'Intelligence Artificielle*, pages 29–32, 4 2021.

[7] Joshua C. Denny and Francis S. Collins. Precision medicine in 2030 – seven ways to transform healthcare. *Cell*, 184(6) :1415–1419, 2021.

[8] William Digan et al. An architecture for genomics analysis in a clinical setting using Galaxy and Docker. *GigaScience*, 6(11), November 2017.

[9] William Digan et al. Evaluating the impact of text duplications on a corpus of more than 600, 000 clinical narratives in a french hospital. In *MEDINFO 2019 : Health and Wellbeing e-Networks for All - Proceedings of the 17th World Congress on Medical and Health Informatics*, volume 264 of *Studies in Health Technology and Informatics*, pages 103–107. IOS Press, 2019.

[10] William Digan et al. Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, 28(3) :504–515, 2021.

[11] Michel Dumontier et al. Bio2RDF Release 3 : A larger, more connected network of Linked Data for the Life Sciences. In *Posters & Demonstrations Track, 13th International Semantic Web Conference, ISWC*, volume 1272 of *CEUR Workshop Proceedings*, pages 401–404. CEUR-WS.org, 2014.

[12] Walid Hafiane et al. Expérimentations autour des architectures d'apprentissage par transfert pour l'extraction de relations biomédicales. In *21ème édition de la conférence "Extraction et Gestion des Connaissances", EGC*, January 2021.

[13] Joël Legrand et al. PGxCorpus, a manually annotated corpus for pharmacogenomics. *Scientific Data*, 7(1) :3, 2020.

[14] Joël Legrand et al. Syntax-based transfer learning for the task of biomedical relation extraction. *Journal of Biomedical Semantics*, 12(1) :16, 2021.

[15] Mehdi Mirzapour et al. French fastcontext : A publicly accessible system for detecting negation, temporality and experiencer in french clinical notes. *Journal of Biomedical Semantics*, 117 :103733, 2021.

[16] Pierre Monnin et al. PGxO and PGxLOD : a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. *BMC Bioinformatics*, 20-S(4) :139 :1–139 :16, 2019.

[17] Pierre Monnin et al. Knowledge-based matching of n-ary tuples. In *Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS*, volume 12277 of *Lecture Notes in Computer Science*, pages 48–56. Springer, 2020.

[18] Pierre Monnin et al. Tackling scalability issues in mining path patterns from knowledge graphs : a preliminary study. In *1st International Conference "Algebras, graphs and ordered sets", ALGOS*, volume 2925 of *CEUR Workshop Proceedings*, pages 123–137. CEUR-WS.org, 2020.

[19] Pierre Monnin et al. Discovering alignment relations with graph convolutional networks : a biomedical case study. *Semantic Web*, pages 1–20, 2021.

[20] Gabin Personeni et al. Discovering associations between adverse drug events using pattern structures and ontologies. *Journal of Biomedical Semantics*, 8(1) :29 :1–29 :13, 2017.

[21] Andon Tchechmedjiev et al. Enhanced functionalities for annotating and indexing clinical text with the NCBO annotator+. *Bioinformatics*, 34(11) :1962–1965, 2018.

[22] Andon Tchechmedjiev et al. SIFR annotator : ontology-based semantic annotation of french biomedical text and clinical notes. *BMC Bioinformatics*, 19(1) :405 :1–405 :26, 2018.

[23] Michelle Whirl-Carrillo et al. Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics*, 92(4) :414, 2012.

[24] Mark D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1) :160018, 2016.