Matching Pharmacogenomic Knowledge: Particularities, Results, and Perspectives[⋆]

Pierre Monnin[1][0000−0002−2017−8426] and Adrien Coulet[2,3][0000−0002−1466−062X]

[1] Orange, Belfort, France
pierre.monnin@orange.com
[2] Inria Paris, Paris, France
[3] Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, Sorbonne Université, Paris, France
adrien.coulet@inria.fr

Abstract. Knowledge in pharmacogenomics (PGx) is scattered across several resources, *e.g.*, reference databases and the biomedical literature. Matching their content would thus lead to a consolidated view of the available PGx knowledge that could, in turn, support multiple downstream applications, including knowledge curation and precision medicine. However, matching atomic units of PGx knowledge is challenging due to their peculiarities: they are of n-ary nature, represented with heterogeneous vocabularies, and with various levels of granularity. In this paper, we frame the matching of PGx knowledge units of various provenance as an instance matching problem. We summarize our work to represent such units within a knowledge graph named PGxLOD, and to match them with a rule-based and a graph embedding-based matching approaches. We then particularly discuss the remaining challenges and how our research artifacts opened to the community could foster new benchmarks and methods for structure-based instance matching.

Keywords: Instance Matching · n-ary Tuple · Preorder · Graph Embedding · Pharmacogenomics.

1 Introduction

The increasing adoption of Linked Open Data (LOD) principles as well as Semantic Web standards and technologies leads to an ever-growing number of resources being published online. Consequently, the knowledge of a domain can be scattered across several complementary, potentially overlapping, resources. That is to say, both similar and complementary knowledge units may be represented across different resources. The aforementioned standards and technologies also allow the concurrent edition of resources by human and software agents. This can lead to duplicates within the same resource. Thus, matching the content of such resources is a first and necessary step towards offering a consolidated view

[⋆] This work was supported by the French National Research Agency (ANR) through the *PractiKPharma* project (ANR-15-CE23-0028 - http://practikpharma.loria.fr).

of the available knowledge of a domain. However, matching similar knowledge units within and across resources requires to face issues such as difference of vocabularies or levels of granularity in the representation of knowledge units.

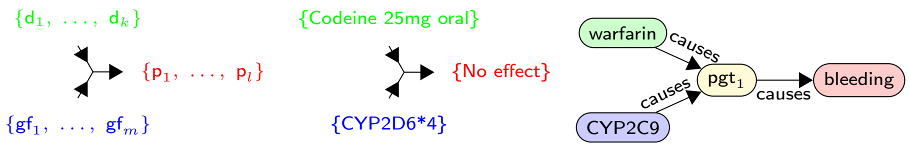
This general observation is also valid in pharmacogenomics (PGx), a domain that studies the influence of genetic factors on drug response phenotypes. Indeed, state-of-the-art knowledge in PGx is mainly scattered across specialized databases such as PharmGKB, and the biomedical literature, *i.e.*, PubMed. Electronic Health Records (EHRs) can also be mined to extract PGx knowledge. Additionally to this scattering, PGx knowledge also suffers from heterogeneous levels of validation. Indeed, some PGx knowledge units have been extensively studied, validated, and are implemented in clinical practice. On the contrary, others have only been observed on reduced cohorts of patients and remain to be further studied and confirmed. Consequently, matching PGx knowledge across resources would, first, offer a consolidated view of the available knowledge of the domain; and second, should ease knowledge curation. Indeed, PharmGKB is manually fed by human curators who continuously review the literature. Connecting similar PGx knowledge units across articles would ease their work by guiding them to sets of relevant articles to consider jointly and confront. This would facilitate the validation, or moderation, of state-of-the-art knowledge.

In this paper, we specify the problem of matching PGx knowledge units of various provenance and its challenges (Section 2). Then, we summarize our research results consisting of a knowledge graph (KG) named PGxLOD, and two matching approaches (Section 3). Finally, we outline new research directions, and advocate for the community reuse of our produced research artifacts (Section 4).

2 Problem Setting

Matching PGx knowledge requires to tackle several specific issues. Indeed, PGx knowledge units are n-ary relationships whose arguments are sets of individuals, *e.g.*, the sets of involved drugs, the sets of involved genetic factors, and the sets of involved phenotypes (Fig. 1a). Such a relationship states that a patient being treated with the specified drugs while having the specified genetic factors will likely experiment the specified phenotypes. To illustrate, Fig. 1b depicts a state-of-the-art PGx relationship. It should be noted that, on the application side, such knowledge units are named *pharmacogenomic relationships*. However, in a mathematical formalism or Semantic Web standards, such units are actually *relation instances* or *relation tuples*. To avoid any ambiguity, we will only refer to PGx knowledge units as *PGx tuples*.

We choose to represent PGx tuples in a KG that follows Semantic Web formalisms to interconnect with ontologies and additional LOD about drugs, genetic factors, and phenotypes. Such interconnections provide additional knowledge that we leverage during the matching process. In such formalisms, only binary predicates exist. Consequently, PGx tuples are reified: tuples become individuals that are linked to their components with binary predicates. For example, in Fig. 1c, the tuple $\texttt{pgt}_1$ is an individual linked to its components with the causes predicate. In such a view, the matching of PGx tuples comes down to an *instance matching* task. Due to the reification of tuples, this process will solely rely on a comparison of neighbors of tuples in the graph, which corresponds to a *structure-based matching* approach.



(a) Abstract relationship (b) Example relationship (c) Reified relationship

Fig. 1. Graphical representation of an abstract (1a), an example (1b), and a reified (1c) pharmacogenomic relationships. The example relationship states that patients having the "*4" version of the $CYP2D6$ gene will not experience the expected effect of codeine. gf stands for genetic factor, d for drug and p for phenotype.

Beside their arity, matching PGx tuples requires to face their incomplete and heterogeneous representations (Fig. 2). These issues inherently lead to *various types of alignments between tuples*, which is somehow unusual in instance matching. For example, tuples can be identical, *e.g.*, tuples on the bottom left. Identical alignments may need to rely on translations and synonyms. Some tuples may be more specific than others, *e.g.*, tuples on the right. Matching such granularity differences may rely on domain-specific orderings such as ontology hierarchies. In addition, a specified argument is more specific than an unknown argument, *e.g.*, vascular disorders is more specific than ??. Tuples may also be related when their arguments are comparable w.r.t. some orderings but not all ordered in the same way (*e.g.*, tuples at the bottom). This type of alignments can also connect tuples whose arguments are somehow related without being strictly comparable.

3 Representing and Matching PGx Knowledge

We developed our own ontology, PGxO, and our own KG, PGxLOD[4] [1], to represent PGx tuples of various provenance. PGxO provides a reduced set of classes and predicates to represent reified PGx tuples. We populated PGxO to create PGxLOD by extracting 50,425 PGx tuples from three main resources: structured data of PharmGKB (3,650 tuples), semi-structured data (namely clinical annotations) of PharmGKB (10,240 tuples), and PubMed abstracts (36,535 tuples). PGxLOD also includes knowledge about drugs, genetic factors, and phenotypes by integrating several LOD graphs and ontologies (*e.g.*, DisGeNET, DrugBank, MeSH). PGxLOD is publicly available and respects LOD and FAIR principles.

[4] https://pgxo.loria.fr - https://pgxlod.loria.fr

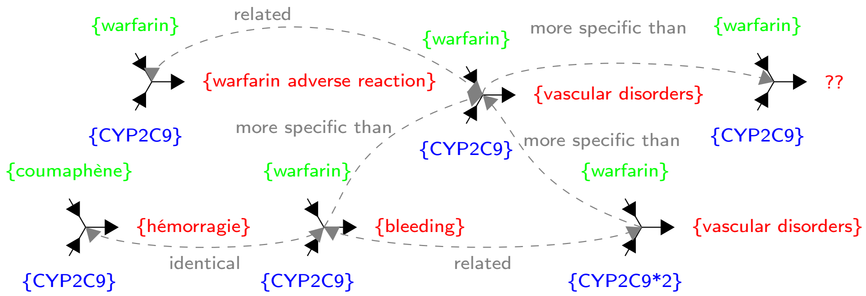


Fig. 2. Example of heterogeneity issues and expected matching results (dashed gray arrows) between PGx tuples. The phenotype is unknown (denoted ??) for one tuple, coumaphène and hémorragie are the French words for warfarin and hemorrhage, which in turn is a synonym of bleeding. CYP2C9*2 is a gene variant that is more specific than the gene CYP2C9 itself, bleeding is more specific than vascular disorder according to an ontology, and vascular disorders is related somehow to warfarin adverse reaction.

To align PGx tuples represented in PGxLOD, we first proposed a symbolic rule-based approach, named tcn3r[5] [2]. In this approach, we see PGx tuples in their mathematical form of n-ary tuples, where each argument is a set of individuals, *e.g.*, from Fig. 1c, $\mathsf{pgt}_1 = (\{\mathsf{warfarin}\}, \{\mathsf{CYP2C9}\}, \{\mathsf{bleeding}\})$. In this view, matching two tuples comes down to comparing their arguments pairwise, before concluding on the relatedness of the two tuples. To this aim, we define two preorders $\preccurlyeq^{\mathsf{P}}$ and $\preccurlyeq^{\mathcal{O}}$ that consider ontology statements, and thus enrich the comparison provided by set operators (*i.e.*, $=$, $\subseteq$). We propose in [2] **five matching rules** that conclude on **five relatedness levels** between tuples. The many resulting alignments illustrate potentialities of our approach and provide insights on PGxLOD. Rule 5 that concludes on the weakest relatedness level generated the most inter-resource alignments, which emphasizes the importance of weaker relatedness levels to align resources and overcome their heterogeneity.

To cope with this need for flexibility, we considered KG embeddings models. Indeed, the continuous aspect of graph embeddings may provide the needed flexibility. We framed our task as a **node clustering task performed on the embedding space**[6] [3]. Consequently, we learn node embeddings using Graph Convolutional Networks and the Soft Nearest Neighbor Loss such that similar PGx tuples have low distances between their embeddings. Then, we apply a clustering algorithm on the node embeddings and consider nodes assigned to the same cluster as similar. To learn node embeddings, we constituted *gold clusters* that are based on the alignments output by our rule-based approach. We showed that integrating domain knowledge by adding inferences in the KG improves clustering performance. We also observed that distances in the embedding space are

[5] https://github.com/pmonnin/tcn3r
[6] https://github.com/pmonnin/gcn-matching

coherent with the "strength" of the different alignments (*e.g.*, smaller distances for equivalences, larger for weak relations). This result corresponds somehow to a rediscovery of KG semantics in the embedding space.

4 Discussion & Perspectives

PGxLOD only contains PGx tuples from the state of the art. Hence, one major perspective resides in developing an "observational" version with results of EHR mining or EHRs themselves. However, such an integration raises several issues related to text mining and data privacy (*e.g.*, anonymization, access control). Beside biomedical research, we think PGxLOD is a useful resource for Computer Science researchers. Indeed, we illustrated with our matching approaches the various and challenging characteristics of PGxLOD: integration of several data sets (*i.e.*, owl:sameAs links) and ontologies (*i.e.*, hierarchies of classes and predicates, predicate inverses and symmetry), and a medium size (*i.e.*, scalability issues). For these reasons, PGxLOD constitutes an interesting real-world KG to experiment matching approaches. That is why, we envision to propose its consideration in the Ontology Alignment Evaluation Initiative.

Both our matching approaches showed how domain knowledge and reasoning mechanisms can serve a structure-based matching. These approaches also present complementary strengths and could foster each other. Two perspectives now lie in *(i)* learning new rules from clusters output by our embedding-based approach and *(ii)* updating alignments between PGx tuples when new tuples are integrated. Other aspects of PGx knowledge and metadata are not currently considered but pave the way for future works. For example, PGxO allows to represent negation within PGx tuples (*e.g.*, drugs not causing a tuple). Consequently, we could match contradictory tuples, which raises several questions such as the geometric representation of contradiction in the embedding space. We could also tackle the task of knowledge validation, *i.e.*, confirming or moderating a knowledge unit based on similar or contradictory units existing in other resources. This would require to leverage alignments and heterogeneous quality metadata such as levels of evidence in PharmGKB, or odd ratios in biomedical articles. Such a knowledge validation approach would, in turn, realize our ultimate objective of offering a consolidated view of PGx knowledge to clinicians.

References

1. Monnin, P., 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)
2. Monnin, P., et al.: Knowledge-based matching of n-ary tuples. In: Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS 2020. LNCS, vol. 12277, pp. 48–56. Springer (2020)
3. Monnin, P., et al.: Discovering alignment relations with Graph Convolutional Networks: A biomedical case study. Semantic Web **13**(3), 379–398 (2022)