Discovering and Comparing Relational Knowledge, the Example of Pharmacogenomics

Pierre Monnin

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
pierre.monnin@loria.fr

Abstract. Pharmacogenomics (PGx) studies the influence of the genome in drug response, with knowledge units of the form of ternary relationships *genomic variation – drug – phenotype*. State-of-the-art PGx knowledge is available in the biomedical literature as well as in specialized knowledge bases. Additionally, Electronic Health Records of hospitals can be mined to discover such knowledge units that can then be compared with the state of the art, in order to confirm or temper relationships lacking validation or clinical counterpart. However, both discovering and comparing PGx relationships face multiple challenges: heterogeneous descriptions of knowledge units (languages, vocabularies and granularities), missing values and importance of the time dimension. In this research, we aim at proposing a framework based on Semantic Web technologies and Formal Concept Analysis to discover, represent and compare PGx knowledge units. We present the first results, consisting of creating an integrated knowledge base of PGx knowledge units from various sources and defining comparison methods, as well as the remaining issues to tackle.

Keywords: Knowledge Discovery · Knowledge Comparison · Semantic Web · Ontology · Formal Concept Analysis · Pharmacogenomics

1 Problem Statement

Pharmacogenomics (PGx) is the study of the influence of genomic variations in the variations in drug response phenotypes. Knowledge in PGx is composed of ternary relationships of the form *genomic variation – drug – phenotype*, where the *phenotype* can be the expected drug outcome or an adverse effect. For example, one well studied PGx relationship is *G6PD:202A – chloroquine – anemia*, stating that patients having the *202A* version of the *G6PD* gene and treated with *chloroquine* will experience *anemia*. PGx is of importance in the implementation of personalized medicine, where drug treatments and drug doses are tailored to the genotype of patients to reduce risks of adverse effects.

State-of-the-art PGx knowledge can be found in *(i)* specialized knowledge bases (*e.g.*, PharmGKB) and *(ii)* the biomedical literature. However, such relationships may have only been observed on reduced cohorts of patients and still remain to be further studied. On the other hand, nowadays, lots of health care data

are digitally available thanks to the use of Electronic Health Records (EHRs). They contain information about diseases, laboratory tests, medical procedures and prescriptions that a patient has experienced. Therefore, mining EHRs to discover PGx knowledge units and then comparing them with the state of the art could confirm or temper poorly validated relationships [7].

However, mining PGx relationships from EHRs and comparing knowledge units from various sources face multiple challenges. First, EHRs data are multivariate, heterogeneous, irregular in time, and sparse [2]. The time dimension should also be carefully taken into account and relationships from a patient level should be generalized to an aggregated level. Then, comparing PGx knowledge units from various sources require to face the heterogeneity of their descriptions in terms of languages, vocabularies and granularities. For example, as genetic data are not common in EHRs, mined relationships will most likely use *proxies*, such as lab measures or specific side effects to a drug treatment that indicate the presence of a specific genomic variation. Knowledge comparison mechanisms should also leverage provenance metadata of knowledge units. Indeed, sources may define quality metrics associated with PGx relationships and representing their "level of validation". Finally, one challenge also resides in dealing with contradictory information coming from different sources. Therefore, in this research, we aim at proposing representation formalisms and discovery and comparison methods in the context of relational and heterogeneous knowledge, formed by PGx relationships.

This paper is organized as follows. Section 2 presents some relevant works w.r.t. the considered challenges. Sections 3 and 4 present the proposed approach as well as the adopted methodology. Section 5 introduces the first results that are discussed in Section 6.

2 State of the Art

To propose a framework for knowledge discovery and comparison in pharmacogenomics, one first needs to define the knowledge representation formalism to adopt. In this direction, one possible solution resides in using Semantic Web technologies, such as OWL and RDF, that allow to represent data and knowledge in a machine-readable format. Such data and knowledge being published and accessible on the Web, it is possible to leverage existing knowledge defined elsewhere when considering a particular data set. Existing works already use Semantic Web technologies to represent Life Sciences data and knowledge. For example, the Bio2RDF project [5] represents and interlinks numerous Life Sciences data sets about drugs, genomic variations, phenotypes, etc. More particularly, ontologies have been created for the PGx domain. However, they are not adapted to the current need of integrating and comparing knowledge units of various sources. For example, the Pharmacogenomic Clinical Decision Support (or Genomics CDS) [19] aims at applying pharmacogenomic guidelines to patient data, to help clinicians in their decisions. Alternatively, the Suggested Ontology for Pharmacogenomics (SO-Phare) [8] focuses on knowledge discovery.

Besides representing PGx knowledge itself, Semantic Web technologies are also used to represent EHRs data. For example, Odgers and Dumontier [18] showed that transforming EHRs data into RDF allowed to connect them with additional knowledge, in the objective of improving knowledge discovery methods. Similarly, Beyan and Decker [4] proposed to use RDF to model temporal relations in EHRs. Indeed, the flexible schema of RDF graphs seems suited to the sparsity and heterogeneousness of EHRs data. Moreover, RDF allows to connect data to existing knowledge repositories. Semantic relations as well as hierarchies of concepts provided by ontologies can also be used to abstract care trajectories of patients.

Mining such abstract care trajectories may be considered as mining sequences of events. Several approaches have already been proposed. One possible approach is to consider Allen's temporal logics to mine temporal patterns, abstracting from exact durations between events [3]. Additionally, lots of health care data are described using concepts of ontologies. For example, drug prescriptions can be encoded using the Anatomical Therapeutic Chemical Classification (ATC). It is also frequent that diseases are encoded with concepts of the International Classification of Diseases (ICD). These ontologies provide hierarchies of concepts, that can be used as background knowledge when abstracting care trajectories. For example, Egho *et al.* [9] proposed to mine heterogeneous multidimensional sequential patterns, taking into account background knowledge represented within ontologies.

Finally, knowledge discovery from EHRs represented as RDF graphs or comparing knowledge in RDF graphs can be achieved using Formal Concept Analysis (FCA) [11]. Indeed, FCA is a mathematical framework grouping objects w.r.t. their common attributes in formal concepts. These formal concepts are organized in a hierarchical structure called a lattice, where the hierarchy indicates a specialization of sets of grouped objects (or dually a generalization of sets of associated attributes). FCA has been extended with Pattern Structures [10] to take into account more complex data than just objects and attributes. FCA and Pattern Structures were already applied on ontology engineering tasks such as mining definitions of ontology concepts [1] and analyzing ontology-based annotations in biomedical documents [6].

3 Proposed Approach

In this research, we propose to develop a framework based on Semantic Web technologies and Formal Concept Analysis to discover, represent, and compare (or *align*) PGx knowledge units from various sources, as presented in Figure 1. It should be noted that the RDF triples considered in this work can be seen as data but also as knowledge.

A first major task resides in building an integrated knowledge base of PGx knowledge units from various sources. To this aim, a common integration schema should be defined. As previously mentioned, existing PGx ontologies are not suited to the current need as they focus on knowledge discovery or reasoning

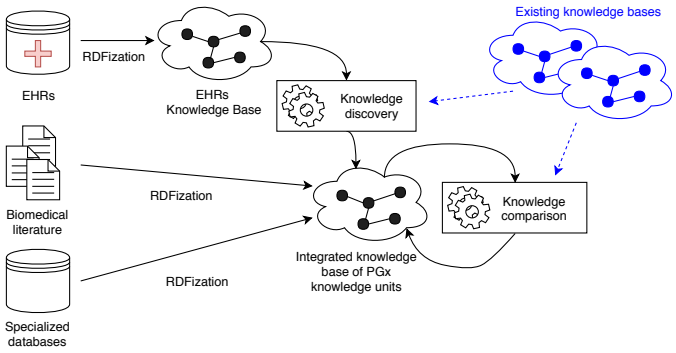

Fig. 1. Proposed framework for knowledge discovery, representation, and comparison in PGx.

instead of integration. This integration schema needs to provide an encoding for provenance metadata. Indeed, the knowledge comparison mechanisms could use information such as quality metrics defined in original sources or by knowledge discovery algorithms (*e.g.*, confidence, support). Finally, provenance metadata could also encode parameters or versions of the developed knowledge discovery algorithms, allowing to compare and evaluate different executions. The flexibility of Semantic Web technologies in defining an integration schema is particularly adapted to our case as PGx knowledge units can be partially discovered from the various considered sources.

By representing PGx knowledge units with Semantic Web technologies, they can be interlinked with knowledge defined elsewhere. This background knowledge can be of particular interest when comparing knowledge units. Indeed, ontologies provide equivalences or subsumption relationships, that could improve identifying equivalent or more specific PGx relationships. For this comparison mechanism, we choose to use Formal Concept Analysis. Indeed, FCA groups similar objects together in formal concepts, which can be used to identify similar PGx relationships. Additionally, the hierarchical structure of the generated lattice can be leveraged to identify relationships more specific or more general than others. Finally, ontologies describing components of PGx relationships and their provided concepts hierarchies can be taken into account by using Pattern Structures.

Finally, we choose to use a two-step approach for knowledge discovery from EHRs. First, EHRs data should be transformed into RDF. The resulting knowledge base can then be mined. As previously mentioned, this RDFization allows to use existing knowledge in the mining algorithms. To mine the RDF representation of EHRs data, FCA can also be considered as patients undergoing similar care trajectories will be grouped into the same formal concepts. Pattern Structures can be used to express sequences representing these care trajectories while benefiting from knowledge defined in ontologies.

4 Methodology

This research work is organized around two main tasks: knowledge discovery from EHRs and knowledge comparison. These two tasks present different methodologies and validation mechanisms.

The methodology for the task of comparing knowledge units can be sketched out as follows:

(1) Define a common integration schema for representing PGx knowledge units from various sources and an encoding for their provenance metadata;
(2) Instantiate this schema with knowledge units from various sources, validating the suitability of the schema to represent these knowledge units and their provenance metadata;
(3) Define and execute comparison methods on the knowledge base resulting from the instantiation process.

It is noteworthy that validating comparison methods will require an expert, to be able to identify whether suggestions of identical, more specific or related relationships are correct. However, we can also develop a first naive comparison method based on domain knowledge rules, constituting a first baseline to be compared with results of advanced methods.

Regarding the task of knowledge discovery from EHRs, as we chose to use a representation using Semantic Web technologies, the methodology should be as follows:

(1) Transform EHRs data into a knowledge base, connect it with existing knowledge defined elsewhere;
(2) Mine the resulting knowledge base for PGx knowledge units;
(3) Validate the discovered knowledge units.

Similarly to the task of comparing knowledge units, the discovery of PGx knowledge units should be validated by a domain expert. However, a first validation step of the discovery algorithms could consist in re-discovering PGx relationships already stated in the biomedical literature from a specific cohort of patients.

5 Results

Learning from the existing ontologies for PGx, we built PGxO, a simple ontology only representing the aspects of PGx needed for integrating and comparing knowledge units from various sources [13]. Based on the W3C Recommendation PROV-O, our work also provides a flexible encoding for provenance metadata, *e.g.*, the source of the knowledge units, quality metrics, *etc.* The ontology and the encoding for provenance metadata were validated by answering *competency questions*. In a first evaluation, we manually instantiated PGxO with knowledge units from *(i)* PharmGKB, *(ii)* the literature and *(iii)* what we though may be discovered in EHRs. In a later evaluation, we instantiated the ontology automatically with knowledge units extracted programmatically from PharmGKB

and the biomedical literature and manually by representing results reported in studies of EHRs [14]. The resulting integrated data set is called PGxLOD.

As a first comparison work, we defined a set of simple *reconciliation rules* [14], identifying when two PGx relationships are referring to the same knowledge unit, if one is a more precise version of the other or if they are related (to some extents). These conclusions are then added to the integrated knowledge base. For example, one rule states that if two PGx relationships involve the same sets of drugs, genomic variations and phenotypes, then they represent the same knowledge unit. Results of executing these reconciliation rules constitute a first baseline to be compared with advanced reconciliation methods.

Finally, regarding comparison methods, we started investigating how FCA could be used to compare knowledge units. In order to compare the class hierarchy of an ontology with the hierarchy formed by a lattice grouping individuals w.r.t. the predicates they are subject of, we defined the notion of *concept annotation* [15, 16]. Each formal concept is annotated with the ontology classes instantiated by all the individuals of the concept. Subsumption axioms are then read from the annotated structure and compared with those already defined in the considered ontology. Using multiple annotations with multiple ontologies, it is also possible to suggest equivalence relationships between classes of different ontologies [17]. Finally, by grouping individuals w.r.t. other individuals they are associated with, it is possible to generalize relationships between individuals to relationships between classes of individuals. This can be seen as a way to describe frequent profiles of predicates, for example families of genes frequently associated with families of drugs.

6 Discussion

Our first results in integrating knowledge units from various sources validate our ontology and the encoding of provenance metadata. The *reconciliation rules* constitute an interesting baseline that will be useful when executing more complex comparison methods. Indeed, these advanced methods should yield the same (or more) comparison results than the reconciliation rules. Additionally, executing the reconciliation rules on PGxLOD led to identify a major shortcoming. As PGx relationships are compared based on the involved drugs, genomic variations and phenotypes, mapping relations identifying equivalent or more precise components are of importance. Therefore, it is important to improve and complete existing mappings, possibly leveraging automatic mappings generated by ontology repositories such as the NCBO Bioportal.

When such mappings are missing, we could leverage unqualified relations between individuals, such as `x-ref` relations. Indeed, they could be used to define or learn similarities between individuals, instead of equivalences. Therefore, one future challenge would be to define and use comparison methods taking into account such similarities, leading to identify similar but not strictly equivalent PGx relationships. Additionally, as PGx knowledge units are ternary relationships, one next challenge resides in using Triadic Analysis [12], an extension

of FCA for ternary relations, with Pattern Structures to integrate background knowledge from ontologies and similarities.

Finally, results of knowledge comparison approaches can be seen as identifying identical knowledge units, more precise ones, and related ones (to some extents). However, we could also envision the occurrence of contradictory knowledge units. In this case, some questions of interest reside in the definition of a contradiction as well as its discovery and its representation.

Acknowledgments

I would like to thank my supervisors Adrien Coulet and Amedeo Napoli and my co-authors Clément Jonquet, Joël Legrand, Mario Lezoche and Andon Tchechmedjiev for our on-going work. This work is supported by the *PractiKPharma* project, founded by the French National Research Agency (ANR) under Grant No. ANR-15-CE23-0028, and by the *Snowball* Inria Associate Team.

References

1. Alam, M., Buzmakov, A., Codocedo, V., Napoli, A.: Mining definitions from RDF annotations using formal concept analysis. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. pp. 823–829 (2015), http://ijcai.org/Abstract/15/121
2. Batal, I.: Temporal data mining for healthcare data. In: Healthcare Data Analytics., pp. 379–402 (2015), http://www.crcnetbase.com/doi/abs/10.1201/b18588-14
3. Batal, I., Fradkin, D., Jr., J.H.H., Moerchen, F., Hauskrecht, M.: Mining recent temporal patterns for event detection in multivariate time series data. In: The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012. pp. 280–288 (2012). https://doi.org/10.1145/2339530.2339578
4. Beyan, O.D., Decker, S.: An RDF based semantic approach to model temporal relations in health records. In: Proceedings of the 9th International Conference Semantic Web Applications and Tools for Life Sciences, Amsterdam, The Netherlands, December 5-8, 2016. (2016), http://ceur-ws.org/Vol-1795/paper6.pdf
5. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2rdf release 2: Improved coverage, interoperability and provenance of life science linked data. In: The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings. pp. 200–212 (2013). https://doi.org/10.1007/978-3-642-38288-8_14
6. Coulet, A., Domenach, F., Kaytoue, M., Napoli, A.: Using pattern structures for analyzing ontology-based annotations of biomedical data. In: Formal Concept Analysis, 11th International Conference, ICFCA 2013, Dresden, Germany, May 21-24, 2013. Proceedings. pp. 76–91 (2013). https://doi.org/10.1007/978-3-642-38317-5_5
7. Coulet, A., Smaïl-Tabbone, M.: Mining electronic health records to validate knowledge in pharmacogenomics. ERCIM News **2016**(104) (2016), http://ercim-news.ercim.eu/en104/special/mining-electronic-health-records-to-validate-knowledge-in-pharmacogenomics

8. Coulet, A., Smaïl-Tabbone, M., Napoli, A., Devignes, M.: Suggested ontology for pharmacogenomics (so-pharm): Modular construction and preliminary testing. In: On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, OTM Confederated International Workshops and Posters, AWeSOMe, CAMS, COMINF, IS, KSinBIT, MIOS-CIAO, MONET, OnToContent, ORM, PerSys, OTM Academy Doctoral Consortium, RDDS, SWWS, and SeBGIS 2006, Montpellier, France, October 29 - November 3, 2006. Proceedings, Part I. pp. 648–657 (2006). https://doi.org/10.1007/11915034_89
9. Egho, E., Raïssi, C., Jay, N., Napoli, A.: Mining heterogeneous multidimensional sequential patterns. In: ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014). pp. 279–284 (2014). https://doi.org/10.3233/978-1-61499-419-0-279
10. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: Conceptual Structures: Broadening the Base, 9th International Conference on Conceptual Structures, ICCS 2001, Stanford, CA, USA, July 30-August 3, 2001, Proceedings. pp. 129–142 (2001). https://doi.org/10.1007/3-540-44583-8_10
11. Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer Science & Business Media (2012)
12. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In: Conceptual Structures: Applications, Implementation and Theory, Third International Conference on Conceptual Structures, ICCS '95, Santa Cruz, California, USA, August 14-18, 1995, Proceedings. pp. 32–43 (1995). https://doi.org/10.1007/3-540-60161-9_27
13. Monnin, P., Jonquet, C., Legrand, J., Napoli, A., Coulet, A.: PGxO: A very lite ontology to reconcile pharmacogenomic knowledge units. PeerJ PrePrints **5**, e3140 (2017). https://doi.org/10.7287/peerj.preprints.3140v1
14. Monnin, P., Legrand, J., Husson, G., Ringot, P., Tchechmedjiev, A., Jonquet, C., Napoli, A., Coulet, A.: PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. bioRxiv p. 390971 (2018). https://doi.org/10.1101/390971
15. Monnin, P., Lezoche, M., Napoli, A., Coulet, A.: Using Formal Concept Analysis for checking the structure of an ontology in LOD: the example of dbpedia. In: Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26-29, 2017, Proceedings. pp. 674–683 (2017). https://doi.org/10.1007/978-3-319-60438-1_66
16. Monnin, P., Napoli, A., Coulet, A.: Discovering subsumption axioms with concept annotation. In: Gestion de Données—Principes, Technologies et Applications (BDA 2017) (2017)
17. Monnin, P., Napoli, A., Coulet, A.: Combining Concept Annotation and Pattern Structures for guiding ontology mapping. In: Proceedings of the 6th International Workshop "What can FCA do for Artificial Intelligence"? co-located with International Joint Conference on Artificial Intelligence and European Conference on Artificial Intelligence (IJCAI/ECAI 2018), Stockholm, Sweden, July 13, 2018. pp. 117–126 (2018), http://ceur-ws.org/Vol-2149/paper11.pdf
18. Odgers, D.J., Dumontier, M.: Mining electronic health records using linked data. AMIA Summits on Translational Science Proceedings **2015**, 217 (2015)
19. Samwald, M., Miñarro-Giménez, J.A., Boyce, R.D., Freimuth, R.R., Adlassnig, K., Dumontier, M.: Pharmacogenomic knowledge representation, reasoning and genome-based clinical decision support based on OWL 2 DL ontologies. BMC Med. Inf. & Decision Making **15**, 12 (2015). https://doi.org/10.1186/s12911-015-0130-1