PGxO: A very lite ontology to reconcile pharmacogenomic knowledge units

Pierre Monnin[1], Clément Jonquet[2,3], Joël Legrand[1], Amedeo Napoli[1] and Adrien Coulet[1]

[1]LORIA (CNRS, Inria Nancy-Grand Est, Université de Lorraine), Vandoeuvre-lès-Nancy, France
[2]LIRMM (CNRS, Université de Montpellier), Montpellier, France
[3]BMIR (Stanford University), Stanford, USA

Introduction

In this article, we present a lightweight and simple ontology named PGxO, that we developed to reconcile and trace knowledge in pharmacogenomics (PGx). PGx studies how genomic variations impact variations in drug response phenotypes [4]. Knowledge in PGx is typically composed of units that have the form of ternary relationships *gene variant–drug–adverse event*, stating that an adverse event may occur for patients having the gene variant when being exposed to the drug. For example, one well studied PGx relationship is *G6PD:202A–chloroquine–anemia*, which states that patients with the version 202A of the gene G6PD and treated with chloroquine (an antimalarial drug) may present an anemia (an abnormally low level of red blood cells in blood). These knowledge units *(i)* are available in reference databases, such as PharmGKB [14], reported in the scientific biomedical literature and *(ii)* may be discovered by mining clinical data such as Electronic Health Records (EHRs). Therefore, knowledge in PGx is heterogeneously described (*i.e.*, with various quality, granularity, vocabulary, etc.). It is also increasing: 40,000 PGx relationships were extracted from the 17,000,000 abstracts available on MedLine in 2008 [2] where there are now 27,000,000 abstracts available. It is consequently worth to extract, then compare, assertions from distinct resources.

We manually developed PGxO, by considering only the essential elements that constitute PGx knowledge units (also referred as PGx relationships in this article), and mapped them to existing ontologies. The formalization of pharmacogenomics concepts and relations within PGxO is not a contribution *per se*, as the ontology mainly reuses content represented in previously defined ontologies. However, PGxO originality is to formalize the ternary relationship previously presented and its provenance thus serving as a global schema for reconciling PGx knowledge units of diverse sources. In particular, we will highlight how we encode the provenance of instances of PGx knowledge units using the PROV Ontology (PROV-O) [9]. Accordingly, two instances extracted from two distinct resources are associated with distinct provenances thanks to PROV-O concepts and relations. We propose a set of rules for reconciling knowledge, *i.e.* for identifying duplicates, or in other terms two instances representing the same knowledge unit. The provenance metadata also allows to associate various quantitative metrics related to PGx relationships such as its level of evidence or the confidence level of the algorithm used to extract the knowledge units. Version and parameters of such algorithms may also be associated, offering the ability to compare outputs of variously tuned executions. By adopting this ontology and defining strict rules for its instantiation, we set up a framework for reconciling, or discerning when necessary, knowledge units reported in, or discovered from various resources. This framework is of importance for the PractiKPharma project, funded by the French National Research Agency, whose goal is to confront PGx knowledge reported in the state of the art (*i.e.*, scientific literature and reference databases) with PGx knowledge discovered from EHRs [3].

Several ontologies have already been developed for pharmacogenomics, but with different purposes, making them inadequate to the present need. In particular, SO-Pharm (Suggested Ontology for Pharmacogenomics) and PO (Pharmacogenomic Ontology) have been developed for knowledge discovery purposes rather than data integration or knowledge reconciliation [1, 6]. The PHARE ontology (for PHArmacogenomic RElationships) has been built for normalizing *gene–drug* and *gene–disease* relationships extracted from texts and is not suitable for representing ternary PGx relationships [2]. More recently, Samwald *et al.* introduced the Pharmacogenomic Clinical Decision Support (or Genomic CDS) ontology, whose main goal is to propose consistent information about pharmacogenomic patient testing to the point of care, to guide physician decisions in clinical practice [12]. We have built PGxO by learning and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3140v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

adapting from these previous experiences. For consistency reasons and good practices, we mapped PGxO concepts to concepts of these four pre-existing ontologies.

Methods

PGxO was developed manually by 3 persons (PM, CJ and AC) in 4 iterations (on June 8th, 2017), following classical ontology construction guidelines [10] and particularly the life cycle of an ontology described in [5]. Accordingly, we achieved the specification, conception, diffusion and evaluation steps of the ontology. In addition, we have connected PGxO to existing ontologies by defining equivalence *mappings*. We also defined *identity rules* which enable to decide when two instances of PGx relationships within the ontology may represent the same knowledge unit.

Specification. The *scope* of PGxO is not to represent all facets of pharmacogenomics, but to represent what we previously defined as PGx knowledge units, *i.e.*, ternary relationships between one (or more) genetic factor, one (or more) drug treatment and one (or more) phenotype; along with their provenance. The *objective* of PGxO is twofold: reconciling and tracing these PGx knowledge units.

Conception. Because of the small size of the ontology, the conception step was realized simultaneously with conceptualization, formalization and implementation steps. PGxO has been implemented in OWL using the Protégé ontology editor.

Diffusion. PGxO was originally shared with collaborators of the PractiKPharma project. It is now publicly available on the BioPortal at `https://bioportal.bioontology.org/ontologies/PGXO`.

Evaluation / Instantiation. For the evaluation, we defined *competency questions* as proposed by Gangemi [7]. These questions are "Does PGxO enable to represent a knowledge unit from the PGx state of the art, along with its provenance?", "Does PGxO enables to represent a knowledge unit of PGxO discovered from clinical data, along with its provenance?", "Does PGxO and its associated rules enable to decide if two knowledge units, with distinct provenances, refer to the same thing?". To answer these questions, we have manually instantiated PGxO with examples of knowledge units (along with their provenance) either from *(i)* the reference database PharmGKB, *(ii)* the literature (extracted by Semantic Medline [11] or FACTA+ [13]), or *(iii)* hand designed according to what may be discovered from EHRs.

Mappings. We manually mapped PGxO concepts to four ontologies related to pharmacogenomics: SO-Pharm, PO, PHARE and Genomic CDS. We also manually completed and incorporated a subset of the mappings automatically computed by the BioPortal in order to connect PGxO with three large spectrum ontologies: MeSH, NCiT and SNOMED CT.

Identity rules. Because the aim of our ontology is to potentially represent multiple provenances for a unique PGx relationship, we defined a set of rules that, when satisfied, enable to decide when two PGx relationships with distinct provenances are in fact referring to the same knowledge unit. Consider two instances of the concept `PharmacogenomicRelationship` r_1 and r_2. We define three concepts for entities associated with r_1 (respectively with r_2): the set of Drugs $D_1 \equiv Drug \sqcap \exists causes.\{r_1\}$ (respectively $D_2 \equiv Drug \sqcap \exists causes.\{r_2\}$), the set of Genetic Factors (which encompasses gene and variant alleles) $G_1 \equiv GeneticFactors \sqcap \exists causes.\{r_1\}$ (respectively $G_2 \equiv GeneticFactors \sqcap \exists causes.\{r_2\}$) and the set of Phenotypes $P_1 \equiv Phenotype \sqcap \exists isCausedBy.\{r_1\}$ (respectively $P_2 \equiv Phenotype \sqcap \exists isCausedBy.\{r_2\}$). `Drug`, `GeneticFactors` and `Phenotype` are three concepts of PGxO. `causes` and `isCausedBy` are two relations of PGxO defined such as $isCausedBy \equiv causes^{-}$. We then define the following rules:

(1) $D_1 \equiv D_2 \sqcap G_1 \equiv G_2 \sqcap P_1 \equiv P_2 \Rightarrow \{r_1\} \equiv \{r_2\}$, *i.e.*, r_1 and r_2 are referring to the same PGx relationship
(2) $D_1 \sqsubseteq D_2 \sqcap G_1 \sqsubseteq G_2 \sqcap P_1 \sqsubseteq P_2 \Rightarrow \{r_1\} \sqsubseteq \{r_2\}$, *i.e.*, r_1 is more specific than r_2
(3) $(D_1 \sqsubseteq D_2 \sqcap G_1 \sqsubseteq G_2 \sqcap P_2 \equiv \bot) \sqcup (D_1 \sqsubseteq D_2 \sqcap G_2 \equiv \bot \sqcap P_1 \sqsubseteq P_2) \sqcup (D_2 \equiv \bot \sqcap G_1 \sqsubseteq G_2 \sqcap P_1 \sqsubseteq P_2) \Rightarrow \{r_1\} \sqsubseteq \{r_2\}$, *i.e.*, r_1 is more specific than r_2.

The three previous rules express the eventuality for r_1 and r_2 to refer to the same PGx relationship or for one to be more specific than the other. In every other situation, we cannot decide if r_1 and r_2, are equivalent or more specific/general.

Results

PGxO consists of 9 concepts, 4 relations and 1 necessary condition (*i.e.*, a subsumption axiom). An overview of PGxO concepts and relations is provided as Supplementary Material (SM1). The ontology is organized around the central concept of `PharmacogenomicRelationship`, whose instances may be

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3140v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

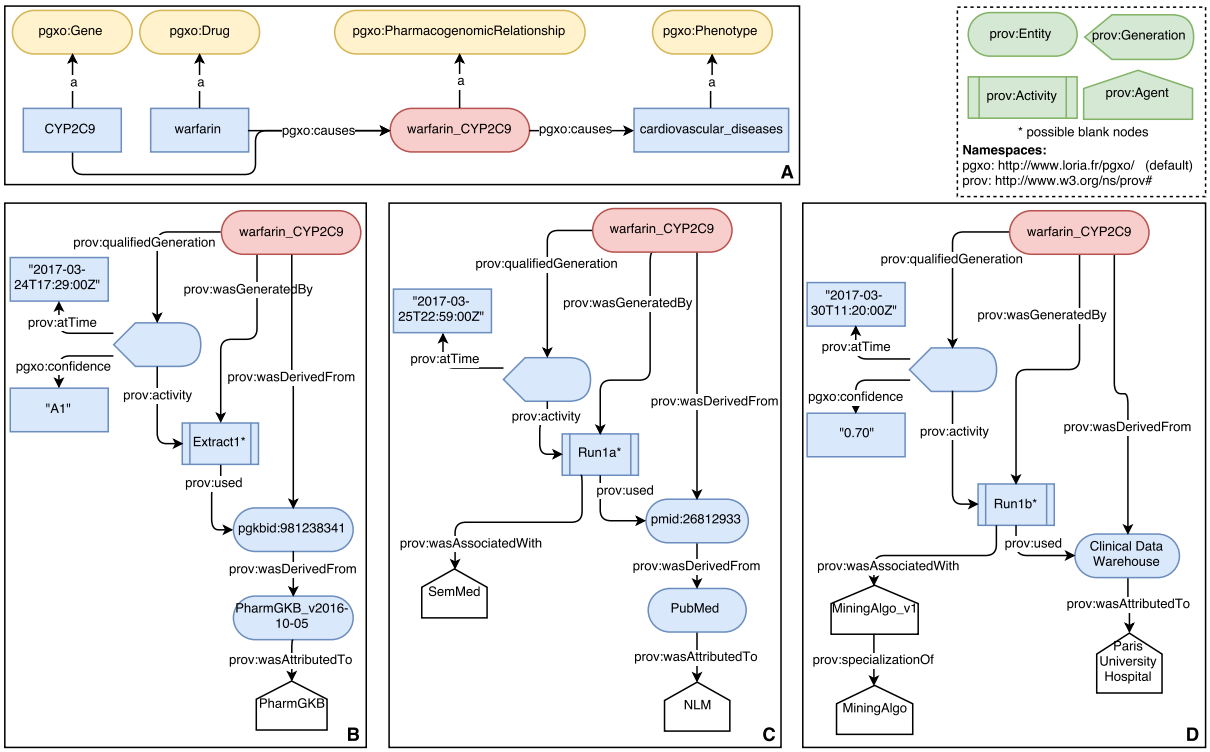

Figure 1: Example of instantiation of PGxO with a relationship (`warfarin_CYP2C9`) and three distinct provenances. Frame **A** represents the pharmacogenomic relationship. Frames **B**, **C** and **D** represent the three distinct provenances, respectively from PharmGKB, literature and EHRs. In these Frames, the shape of the nodes refers to the type of PROV-O concepts they are instance of. The different types of PROV-O concepts and their associated shapes are listed in the upper right caption. Numeric IDs used in Frames B and C respectively correspond to the PharmGKB annotation identifier and to the PubMed identifier used to extract the PGx relationship.

caused by a `GeneticFactor` (such as a gene or gene variants), a `Drug` and causes a `Phenotype`. For the evaluation of PGxO, we tried answering competency questions by instantiating it with two PGx relationships (*warfarin_CYP2C9* and *warfarin_CYP2C9_1*). The first relationship has three distinct provenances and the second one is a specialization of the first one, with another distinct provenance. Figure 1 represents an extract of PGxO with the instantiation of the *warfarin_CYP2C9* relationship (Frame A) and its three distinct provenances (Frames B, C and D).

First, *CYP2C9_warfarin* was found in PharmGKB with a confidence level “A1”. This provenance is expressed by instantiating PROV-O concepts and relationships as illustrated in Frame B. Second, another PGx relationship was found in SemMed with the same associated drug, gene and phenotype. In this particular case, rule (1) enables to decide that this new instance is equivalent to *CYP2C9_warfarin*. Accordingly, we associate *CYP2C9_warfarin* with this new provenance as illustrated in Frame C. Finally, Frame D shows a possible example of clinical data provenance for the same instance. It illustrates the ability of our approach to capture such example, even though we have not run any mining algorithm on a clinical data warehouse yet. A similar PGx relationship was also extracted by FACTA+, but with a more precise adverse event as phenotype. Indeed, FACTA+ proposes a relationship causing *heart_block*. *Heart_block* (D006327) is more specific than *cardiovascular_diseases* (D002318) according to MeSH, therefore rule (2) allows us to identify that we need a new instance of `PharmacogenomicRelationship`, named here *CYP2C9_warfarin_1*, to associate this new provenance with. This new instance was specified as more specific than *CYP2C9_ warfarin* (*i.e.*, $\{CYP2C9\_warfarin\_1\} \sqsubseteq \{CYP2C9\_warfarin\}$). The full instantiation is provided in an OWL file (SM2) and presented in a global figure representing the instantiated ontology (SM3).

In our examples, PGx relationships are expressed using genes. However, our model offers to instantiate PGx relationships involving genomic variations (*e.g.*, variants, haplotypes). Mappings from PGxO

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3140v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

to other PGx ontologies are provided in Supplementary Material SM4, and mappings to large spectrum ontologies in SM5. Among the 9 concepts of PGxO, we were able to map 3 to SO-Pharm, 7 to PHARE, 2 to PO, 2 to Genomic CDS, 7 to MeSH, 7 to NCIt and 6 to SNOMED CT.

Conclusions

In this paper, we presented PGxO, a lightweight ontology for pharmacogenomics, and more importantly, we proposed a set of rules for its instantiation. Using PGxO, one can represent multiple provenances for pharmacogenomic knowledge units, and reconcile duplicates when they come from distinct sources. Thanks to the provided mappings, more expressive ontologies can be leveraged to use the reconciled knowledge in further applications. Because our ontology is minimal and the set of rules reduced, our ontology is easy to understand, adapt and reuse. In the future, we will represent our rules with the SWRL standard [8] and include them in the ontology. Also, we plan to leverage on Semantic Web standards to connect components of PGx knowledge units (*e.g.*, drugs, genes, phenotypes) with Linked Open Data entities elsewhere defined. The main use case of PGxO is to be instantiated by various software agents extracting PGx knowledge from different sources. The resulting knowledge base that we aim at populating within the PractiKPharma project will serve as a framework for confirming or tempering state of the art knowledge in PGx. To this end, our identity rules constitute a very first step.

Supplementary Material

Supplementary material is available at `https://github.com/practikpharma/PGxO`. It includes:

- SM1: Figure presenting PGxO concepts and relations, `./blob/master/doc/pgxo-overview.pdf`
- SM2: An instantiated version of PGxO, `./blob/master/doc/pgxo_with_instances.owl`
- SM3: Global figure presenting the instantiation of PGxO, `./blob/master/doc/global-fig.pdf`
- SM4: Mappings from PGxO to four ontologies related to PGx, `./blob/master/doc/mapp1.owl`
- SM5: Mappings from PGxO to three large spectrum ontologies, `./blob/master/doc/mapp2.owl`

References

[1] Adrien Coulet et al. Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing. In *OTM 2006 Workshops*, pages 648–657, 2006.

[2] Adrien Coulet et al. Integration and publication of heterogeneous text-mined relationships on the Semantic Web. *Journal of Biomedical Semantics*, 2(S-2):S10, 2011.

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

[4] Leslie Dickmann and Joseph Ware. Pharmacogenomics in the age of personalized medicine. *Drug discovery today. Technologies*, Sep - Dec;21-22:11–16, 2016.

[5] Rose Dieng, Olivier Corby, Alain Giboin, and Myriam Ribiere. Methods and tools for corporate knowledge management. *International journal of human-computer studies*, 51(3):567–598, 1999.

[6] Michel Dumontier and Natalia Villanueva-Rosales. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics*, 10(2):153–163, 2009.

[7] Aldo Gangemi. Ontology design patterns for semantic web content. In *ISWC 2005*, pages 262–276, 2005.

[8] Ian Horrocks et al. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member submission*, 21:79, 2004.

[9] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, et al. PROV-O: The PROV Ontology. *W3C recommendation*, 30, 2013.

[10] Natalya F Noy et al. Ontology development 101: A guide to creating your first ontology, 2001.

[11] Thomas C. Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, et al. Semantic MEDLINE: an advanced information management application for biomedicine. *Inf. Services and Use*, 31(1-2):15–21, 2011.

[12] Matthias Samwald et al. Pharmacogenomic knowledge representation, reasoning and genome-based clinical decision support based on OWL 2 DL ontologies. *BMC Medical Informatics & Decision Making*, 15:12, 2015.

[13] Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun'ichi Tsujii, and Sophia Ananiadou. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, 27(13):111–119, 2011.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3140v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017