PGxCorpus and PGxLOD: two shared resources for knowledge management in pharmacogenomics

Pierre Monnin, Joël Legrand, Adrien Coulet

PGxCorpus and PGxLOD: two shared resources for knowledge management in pharmacogenomics

Pierre Monnin[1,2], Joël Legrand[1,3] and Adrien Coulet[1,4,5]
[1] LORIA, Université de Lorraine, CNRS, Inria, Nancy, France
[2] Orange, Belfort, France
[3] CentraleSupélec, Metz, France
[4] Centre de Recherche des Cordeliers, Inserm, Univ. Paris Cité, Sorbonne Univ., Paris, France
[5] Inria Paris, Paris, France

Corresponding author: adrien.coulet@inria.fr

Pharmacogenomics (PGx) studies the impact of genetic factors on drug response phenotypes. Atomic knowledge units in PGx have the form of ternary relationships linking one or more drugs, one or more genetic factors, and one or more phenotypes. Such relationships state that a patient having the specified genetic factors and being treated with the specified drugs is likely to experience the given phenotypes. PGx knowledge is of particular interest for the development of precision medicine which aims at tailoring drug treatments to each patient to reduce adverse effects and maximize drug efficacy. However, PGx knowledge is scattered across many sources (*e.g.*, reference databases, the biomedical literature) and suffers from very heterogeneous levels of validation, *i.e.*, some PGx relationships are extensively studied and have been translated into clinical practice, but most are only observed on small-size cohorts or not reproduced yet and necessitate further investigation. Consequently, there is a strong interest in extracting and integrating knowledge units from these different sources into a unique place to provide a consolidated view of the state-of-the-art knowledge of this domain and drive to the validation, or moderation, of insufficiently validated knowledge units. To this aim, we created and share with the community two resources: PGxCorpus and PGxLOD.

PGxCorpus is a manually annotated corpus, designed for the automatic extraction of PGx relationships from text [1]. In specialized domains such as PGx, state-of-the-art approaches rely on supervised models trained or fine-tuned on previously annotated texts. PGxCorpus has been built by 11 annotators and consists of 945 sentences from PubMed abstracts annotated with *(i)* PGx entities of interest, *i.e.*, genetic factors (*e.g.*, genes, variants, haplotypes), drugs, and phenotypes, and *(ii)* relationships between these entities, associated with a type (*e.g.*, causes, decreases, transports) and an attribute (positive, hypothetical, or negative). It includes 2,875 relationships, each seen at least four times and in total by four different annotators. PGxCorpus is available at http://pgxcorpus.loria.fr.

PGxLOD is a knowledge graph that gathers 50,435 PGx relationships extracted from expert databases such as PharmGKB and from the literature [2]. It implements Semantic Web and FAIR best practices. Relationships of the literature are extracted with a model trained on PGxCorpus. Besides PGx relationships, PGxLOD includes knowledge about genetic factors, drugs, and phenotypes (*i.e.*, PGx key entities) imported from ClinVar, DisGeNET, DrugBank, SIDER, etc. to compose a graph of 88 million triples. We have paid particular attention at connecting PGx relationships that come from independent data sources but may be similar or equivalent with the development of both rule-based and machine learning matching approaches. PGxLOD is available at https://pgxlod.loria.fr.

These resources open perspectives with applications such as predicting pharmacogenes or mining molecular explanations of adverse drug reactions [3]. Additional analyses of PGxLOD offer the potential to guide PGx research by identifying knowledge that requires additional validation. Besides biomedical applications, PGxCorpus and PGxLOD offer challenging experimental settings to both NLP (discontinued entities, ternary relationships) and graph mining tasks (heterogeneity and arity of PGx relationships).

References

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

[2] P. 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.

[3] E. 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.