Suggesting valid pharmacogenes by mining linked data

Kevin Dalleau, Ndeye Coumba Ndiaye, Adrien Coulet

Suggesting valid pharmacogenes by mining linked data

Kevin Dalleau[1], Ndeye Coumba Ndiaye[2], and Adrien Coulet[1]

[1] LORIA (CNRS, Inria NGE, University of Lorraine), Nancy, France
[2] UMR U1122 IGE-PCV (INSERM, University of Lorraine), Nancy, France
{kevin.dalleau@gmail.com, ndeye-coumba.ndiaye@univ-lorraine.fr, adrien.coulet@loria.fr}

Abstract. A standard task in pharmacogenomics research is identifying genes that may be involved in drug response variability, *i.e.*, pharmacogenes. Because genomic experiments tended to generate many false positives, computational approaches based on the use of background knowledge have been proposed. Until now, those have used only molecular networks or the biomedical literature. Here we propose a novel method that consumes an eclectic set of linked data sources to help validating uncertain drug–gene relationships. One of the advantages relies on that linked data are implemented in a standard framework that facilitates the joint use of various sources, making easy the consideration of features of various origins. Consequently, we propose an initial selection of linked data sources relevant to pharmacogenomics. We formatted these data to train a random forest algorithm, producing a model that enables classifying drug–gene pairs as related or not, thus confirming the validity of candidate pharmacogenes. Our model achieve the performance of F-measure=0.92, on a 100 folds cross-validation. A list of top candidates is provided and their obtention is discussed.

1 Introduction

Pharmacogenomics (PGx) studies how individual gene variations cause variability in drug responses [36]. A state of the art of PGx is available and constitutes a basis for implementing personalized medicine, *i.e.*, a medicine tailored to each patient by considering in particular her/his genomic context. This state of the art lies both in the biomedical literature and in specialized databases [15, 35], but a large part of it is controversial, and not yet applicable to medicine. Indeed, its results from studies difficult to reproduce and that do not fulfil statistical validation standards for two main reasons: the small size of populations involved in studies because of the rarity of gene variants studied and the potential coaction of several variants [24, 37]. It is consequently of interest to the PGx community to explore any source of evidence that may contribute to confirming or moderating PGx state of the art. So far, existing works used either molecular network databases or the biomedical literature (see Section 2 for references).

Linked Open Data (LOD) are constituting a large and growing collection of datasets represented in a standard format (that includes the use of RDF and URIs), partially connected to each other and to domain knowledge represented within semantic web ontologies [5]. For these reasons, LOD offer novel opportunities for the development of

successful data integration and knowledge discovery approaches. The recent availability of LOD is particularly beneficial to the life sciences, where relevant data are spread over various data sources with no agreement on a unique representation of biological entities [2]. Consequently, data integration is an initial challenge one faces if one wants to mine life science data considering several data sources. Various initiatives such as Bio2RDF, the EBI platform, PDBj and Linked Open Drug Data (LODD) aim at pushing life sciences data into the LOD cloud with the idea of facilitating their integration [7, 13, 25, 31]. It results from these initiatives a large collection of life-science data unequally connected but in a standard format and available for mining. Despite good will and emerging standard practices for publishing data as LOD, several drawbacks make their use still challenging [17, 28]. Among existing difficulties we can cite the limited amount of links between datasets and the limits of implementations of federated queries.

In this paper we present a novel method that consists in mining an eclectic set of linked data sources to help validating uncertain drug–gene relationships. This method can be divided in three steps: *first*, selecting, connecting and, when necessary, publishing relevant PGx linked data; *second*, formatting linked data in a set of instances, suitable to train a machine learning algorithm; *third*, training a random forest model, subsequently used to classify and rank candidate pharmacogenes.

The paper is organized as follow: next section introduces some related works; section 3 presents successively the preparation, formatting and mining of PGx linked data; section 4 provides the results of training and classification tasks; the last section discusses the proposed method and its results.

2 State of the Art

2.1 Pharmacogenomics data and linked data

PharmGKB is a comprehensive database about PGx that includes manually annotated gene–drug relationships [35]. Recently, PharmGKB distinguished well validated gene–drug relationships from insufficiently validated ones, pointing at knowledge in need of additional validation [30]. Parts of PharmGKB have been transformed and published in RDF by the Bio2RDF project, enabling SPARQL queries [4]. Several other databases provides data that are indirectly relevant to PGx, such as DrugBank (providing for instance drug–target relationships), CTD, Sider, OMIM. Hoehndorf *et al.* integrated and made available a set of PGx related data that includes PharmGKB, DrugBank and CTD, using semantic web technologies [21]. They used the integrated dataset to identify pathways that may be perturbed in PGx. In this effort of publishing PGx data, Coulet *et al.* extracted about $40,000$ PGx relationships from the biomedical literature and published them in the form of RDF statements [9].

2.2 Mining linked data

Suggesting novel drug–gene relationships from an RDF graph can be described as a link prediction problem. Many works have focused on the link prediction problem studying various approaches such as machine learning [3, 22], graph mining [33], identity resolution [6, 34] and data visualisation [20]. Some of these methods obtain good results,

but all are dependent from the quality of input graphs and are hard to reuse for new applications. In relation with PGx research, Percha *et al.* mined with a Random Forest (RF) algorithm the set of RDF statements extracted from text by Coulet *et al.* and predicted successively drug–drug interactions [29].

2.3 Discovery of Pharmacogenes

Hansen *et al.* proposed a method based on a logistic classifier to generate candidate pharmacogenes, using data from PharmGKB, DrugBank, and protein–protein interactions from InWeb [19]. An issue here is that PharmGKB and DrugBank are manually curated from the literature and are consequently expensive to maintain and update. Garten *et al.* answered this issue by proposing an automatic method that consider directly (and only) the literature [16]. They improved the results obtained by Hansen *et al.* by considering gene–drug pairs co-occurring in sentences of the PGx literature. Recently, Funk *et al.* proposed also to use the biomedical literature, plus GO annotations, to identify pharmacogenes [14]. They achieve high F-measure and AUC (0.86 and 0.86), but proposed a binary classification that avoids any ranking of the candidates.
Semantic web technologies have also been experimented for PGx knowledge discovery. Dumontier and Villanueva-Rosales proposed a knowledge representation of the domain and benefit from reasoning mechanisms to answer sophisticated queries related to depression drugs [12]. Coulet *et al.* used patient data to instantiate a DL knowledge base, then extracted association rules from it to identify *gene variant–drug response* associations [8]. More generally, advantages that semantic web technologies may offer to PGx and personalized medicine are listed in [32].

In this paper, we report about a selection of data source that we think relevant to mine for validating pharmacogenes. When necessary, we transformed in RDF graphs and interconnected these data, respecting semantic web precepts to facilitate later addition/removal of sources. Finally, we formatted obtained RDF data to train a RF model, subsequently used to classify candidate pharmacogenes.

3 Methods

3.1 Preparation of the PGx linked data

Data selection Initial step is to select a set of data that include relevant features about PGx drug–gene relationships. Figure 1 gives an overview of the type of data we consider in this study: data about three types of entities (*Gene*, *Phenotype* and *Drug*) and relationships between them (*i.e.*, *Gene–Phenotype*, *Phenotype–Drug* and *Gene–Drug* relationships). To obtain these data, we selected sources manually but oriented our selection to sources providing typed relationships and limited ourselves to two sources per relationship. As a result, we selected ClinVar and DisGeNET for Gene–Phenotype; SIDER and Medi-Span for Phenotype–Drug; DrugBank for Gene–Drug relationships. PharmGKB completes the set of data sources to enable building the training and test sets (see subsection 3.3).

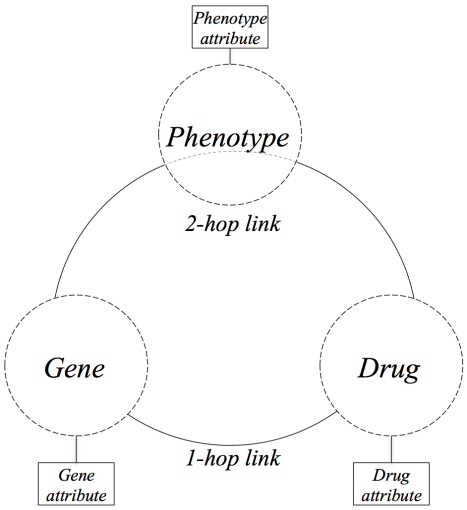

Fig. 1. General view of the type of entities and relationships considered in this study. Naming of different parts (*i.e.*, *1-hop* , *2-hop links* and *gene, phenotype* and *drug attributes*) is used later, in the step of formatting of the linked data (see subsection 3.2).

Data RDFization The second step is about turning selected data in standardized RDF graph. We benefit from the fact that DisGeNET[3], SIDER[4] and DrugBank[5] are already available online in the form of LOD and reused them. We completed the Bio2RDF version of PharmGKB locally with gene–drug relationships manually annotated by PharmGKB but not openly distributed [4]. Similarly, we transformed drug indications and side-effects from Medi-Span in the form of RDF triples and loaded them into our local SPARQL server. DisGeNET includes data from ClinVar, but because it includes only a part of it, we made our own RDF version of ClinVar[6] following guidelines and scripts of the Bio2RDF project. This last dataset will be made available on the Bio2RDF portal soon. Figure 2 presents the detailed schema (*i.e.*, type of entities and relationships) of the linked data we consider for mining.

Mapping definition The third step of the data preparation is to define mappings between entities of the same type but of various provenance. Figure 2 illustrates that for each type of entity (Gene, Phenotype or Drug), data we consider may come from four distinct sources. To define mappings, we first relied on standard identifiers such as NCBI Gene ID found in DisGeNET and ClinVar URIs and UMLS CUI found in DisGeNET, ClinVar, SIDER and Medi-Span. We defined regular expressions over URIs to isolate identifiers and when two match, we define a mapping. Figure 3 shows two entities, `clinvar:1956` and `disgenet:1956` that share a unique identifier within different namespaces.

[3] DisGeNET endpoint: `http://rdf.disgenet.org/sparql/`
[4] SIDER endpoint: `http://sider.bio2rdf.org/sparql`
[5] DrugBank endpoint: `http://drugbank.bio2rdf.org/sparql`
[6] ClinVar in RDF: `http://dbs.kevindalleau.fr/sparql`

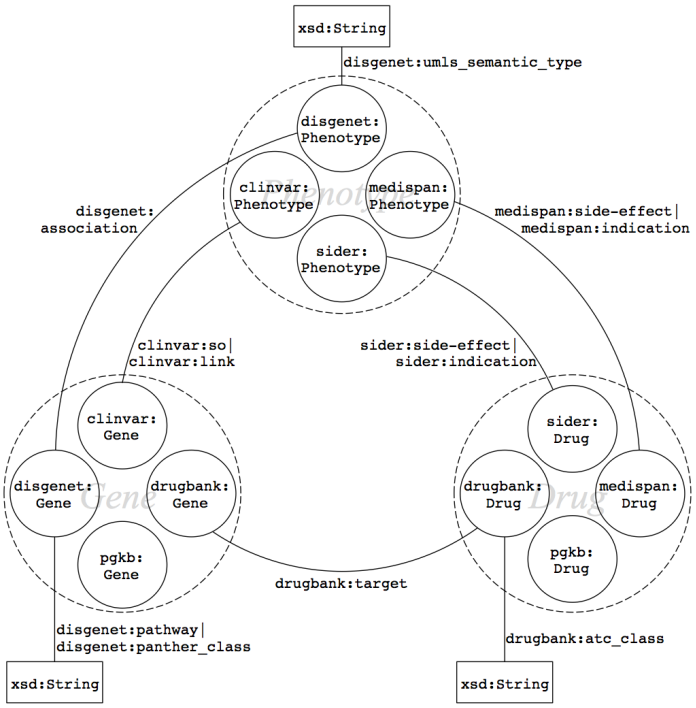

Fig. 2. Schema of the data selection made for this study. Entities are of three distinct types: Gene, Phenotype and Drug. Gene–Phenotype relationships are coming from ClinVar and DisGeNET, Phenotype–Drug relationships from SIDER and Medispan, Gene–Drug relationships from DrugBank. In addition, we included gene and drug entities from PharmGKB to enable building the training and test sets (detailed in subsection 3.3). Equivalence mappings are defined between entities of the same type but of different provenance.

Second, when no standard identifier exists, we relied on services provided by `biodb.jp` to obtain cross-references between identifiers and, accordingly, define mappings [23]. We implemented a tool named biojp2rdf that transforms in RDF the cross-references provided by `biodb.jp` [10]. We loaded these data into a SPARQL server dedicated to the resolution of identity between entities of the same type.

For the sake of computation of SPARQL queries, of independence from third party RDF platforms and of license of certain data, we loaded all considered data on our local server, using Apache Jena Fuseki.

3.2 Formatting the PGx linked data

Linked data are in the form of graphs, whereas machine learning algorithms, such as RF take as an input a feature matrix. Consequently, we needed to format our PGx linked data in the form of such a matrix. Each line of a feature matrix represents an instance and each column represents a feature describing the instances. In this work, we hypothesize

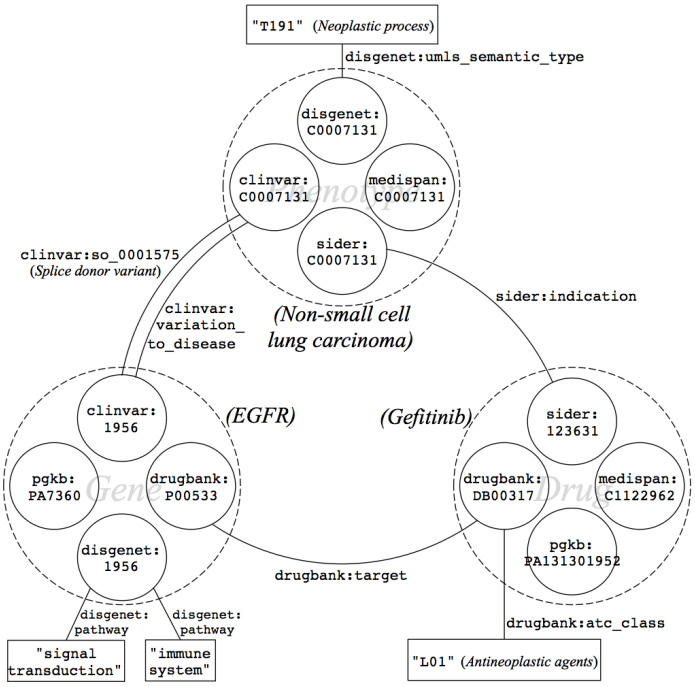

Fig. 3. Sample of PGx linked data, describing relationships between *EGFR* gene, *Gefitinib* and *Non-small cell lung carcinoma*. Entities of same type but of different provenance are mapped to each other.

that paths that exist between 2 entities in linked data may describe their relationship. Consequently we aimed at encoding, in the matrix, paths between genes and drugs. To contain the size of the matrix, we considered only paths of length 1 and 2, hereafter named *1-hop* and *2-hop links*. In addition to links, we encoded few *attributes* that qualify the drug, the gene and the potential intermediate phenotypes (see Figure 1). As a result, each instance is corresponding to a combination of paths between a drug and a gene, plus a combination of potential attributes. One drug–gene pair may be described in the matrix by several instances, but each instance describes a unique pair. The maximal number of instances that may describe the relationship between drug d and gene g is

$$n_{d\text{-}g} = |\{1\text{-}hop\ links\}_{d\text{-}g}| \times |\{2\text{-}hop\ links\}_{d\text{-}g}| \times |\{att\}_d| \times |\{att\}_g| \times \prod_{i=1}^{m} |\{att\}_{p_i}|$$

where $|\{\text{x-}hop\ links\}_{d\text{-}g}|$ is the number of distinct x-hop links between g and d, $|\{att\}_y|$ is the number of distinct attributes of the entity y and p_i are intermediate phenotypes of 2-hop links. The various amount of available data for distinct d-g pairs explains that $n_{d\text{-}g}$ may be very different (from one to several thoushands) from one pair to another. Table 1 shows an example of matrix obtained when formatting the sample of linked data represented in Figure 3.

Table 1. Example of a feature matrix generated from linked data represented in Figure 3. All the instances (*e.g.*, lines) describe the same drug–gene relationships (EGFR–Gefitinib), which is annotated as associated in PharmGKB (*Class*=1).

ID	*Gene attribute*	*Phenotype attribute*	*Drug attribute*	*1-hop link*	*2-hop link1*	*2-hop link2*	*Class*
PA7360-PA131301952	signal transduction	T191	L01	drugbank:target	clinvar:so_0001575	sider:indication	1
PA7360-PA131301952	immune system	T191	L01	drugbank:target	clinvar:so_0001575	sider:indication	1
PA7360-PA131301952	signal transduction	T191	L01	drugbank:target	clinvar:variation_to_disease	sider:indication	1
PA7360-PA131301952	immune system	T191	L01	drugbank:target	clinvar:variation_to_disease	sider:indication	1

3.3 Mining the PGx linked data

Training set To classify drug–gene pairs as associated or not with regards to PGx data, we use Random Forest (RF) that is a supervised machine learning algorithm. RF requires for training two sets of instances: positives and negatives. Our sets of positives and negatives are drug–gene pairs annotated as associated or not according to PharmGKB (version of June 1^{st}, 2013) that are related by at least one 2-hop link and that are associated with a high level of validation in PharmGKB (*i.e.*, level=1 or 2) [30]. PharmGKB includes $2,542$ positive and 373 negative relationships, only 78 and 8 of those have a 2-hop link and 51 and 8 of those have a high level of validation.
To balance the number of positives and negatives, we enriched the set of negatives with 43 drug–gene pairs that are generated randomly, but checked to be absent from DGIdb (the Drug Gene Interaction database), which collects gene–drug relationships from various sources [18]. Resulting 51 positive pairs and 51 negative pairs are used as seeds in our formatting approach to explore PGx linked data and generate respectively $4,618$ and $1,170$ instances[7]. Note that balancing the number of positive and negative pairs results in unbalancing the number of positive and negative instances.

Test set We considered $1,760$ drug–gene pairs insufficiently validated, *i.e.*, associated with the evidence level 3 or 4 in PharmGKB. From them, we kept the 82 that have at least one 2-hop link. These pairs served as seeds to constitute the $13,500$ instances of our test set, according to our formatting approach. 3.2.

Multi-instance classification and candidate ranking With RF prediction, a probability distribution value may be used to evaluate the confidence of the model for classifying a new instance and then rank classified instances. However, the drug–gene pairs that we classify are typical examples of multi-instance objects, also named bag of instances, since they are not represented by a single instance but by several ones. Even if RF may

[7] The training set is open at `http://www.loria.fr/~coulet/training_set.csv`

be trained on single instances to classify bags of instances [26], RF output associates a class and a probability to an instance, not to a bag of instances. Then, we required additional treatments to classify and rank bags of instances. We consider a pair as associated only if all the instances of a bag are classified as positive. For the probability, we compute the arithmetic mean $\bar{p}$ of probabilities of instances of the bag [1].

4 Training and Classification Results

We trained and evaluated our model using the Weka implementation of the RF and 100-fold cross validation. Its processing lasts 13 seconds on a Intel i5-4570 (3,2 GHz). Table 2 presents the results of this evaluation.

Table 2. Results of the 100-fold cross validation of our RF model, trained on PharmGKB well validated data.

Class	*Precision*	*Recall*	*F-Measure*
1 *(positive)*	0.936	0.969	0.952
0 *(negative)*	0.858	0.738	0.794
Weighted Average	0.92	0.922	0.92

The $13,500$ instances of our test set have been classified either as positive or negative in about 5 seconds. The top-10 pairs predicated as positive according to the instance classification is provided in Table 3.

Table 3. 10-Top candidates of drug–gene pairs predicted from our PGx linked data.

Rank	*Gene*	*Drug*	$\bar{p}$	*Rank*	*Gene*	*Drug*	$\bar{p}$
1	APOE	Rosuvastatin	0.874	6	SLC10A1	Thalidomide	0.860
2	APOE	Fenofibrate	0.874	7	ABCC6	Thalidomide	0.853
3	APOE	Simvastatin	0.874	8	HLA-B	Carbamazepine	0.851
4	ACAPG	Vincristine	0.864	9	HLA-B	Lamotrigine	0.851
5	AADRB2	Risperidone	0.862	10	HLA-B	Oxcarbazepine	0.851

5 Discussion and Conclusion

The aim of this work is to study the usefulness of LOD mining for PGx by setting up a baseline experiment. The method we propose is simple and may be improved. We list here some potential improvements. First, the size of our training set is small because we aimed at containing the amount of missing data by considering only pairs of entities related by a 2-hop link. We may enlarge the training set by removing this constraint.Second, we propose a selection of data sources that was made arbitrary by the authors, whereas attribute selection methods may guide this process. Indeed, our choice for linked data framework is motivated by the fact that we want to ease the addition/removal of data sources for enabling the selection of best features out of many sources. A preliminary processing of the information gain (InfoGain) on our features

shows that the type of the 1 hop-link obtained from DrugBank, which may have the value `drugbank:target` or `n/a` is useless on its own (InfoGain=4.10^{-16}), whereas gene and drug attributes are of importance (InfoGain=0.22 and 0.12). Third, the formatting of RDF graph data in the form of a feature matrix may be improved. In our case, a drug–gene pair from the training or test set is encoded by multiple instances, whereas RF classifies single instances [1]. To classify multi-instance objects, we may consider more sophisticated methods such as *MIForests*, a multi-instance learning algorithm for randomized trees [26]. Fourth, the choice for RF algorithm may be discussed and we may compare its results with alternative machine learning algorithms such as SVM that have been successfully used on LOD. This will require to adapt the formatting of the RDF graph data to comply with the chosen algorithm [11, 27].
A clear limitation of our study is the coarse grain of entities we considered. State of the art in PGx reports about relationships between *genomic variant*, sometimes *haplotypes*, and *drug response phenotype*, whereas we are considering simply genes and phenotypes. This is harmful since a unique gene may host two variants, one that impacts drug response and one that does not.

We presented in this paper a method to help validating candidate pharmacogenes using linked data, and its initial results. We selected, published and interconnect data relevant to PGx domain in the form of RDF graphs. Then we formatted these data to train a RF classifier. We used this classifier to identify and rank candidate pharmacogenes. Potential improvements to our method have been identified, however this baseline experiment present promising results, achieving a F-measure=0.92. Top candidate pharmacogenes underlined by the methods will be investigated to evaluate how systematically we can confirm or moderate insufficiently validated PGX knowledge.

References

[1] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[2] Erick Antezana, Martin Kuiper, and Vladimir Mironov. Biological knowledge management: the emerging role of the Semantic Web technologies. *Briefings in Bioinformatics*, 2009.

[3] Veli Bicer, Thanh Tran, and Anna Gossen. Relational kernel machines for learning from graph-structured RDF data. In *ESWC*, pages 47–62, 2011.

[4] Bio2RDF project. PharmGKB endpoint [visited Sept. 2, 2015]: `http://cu.pharmgkb.bio2rdf.org/sparql`.

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[6] C. Y. A. Brenninkmeijer, I. Dunlop, C. A. Goble, A. J. G. Gray, S. Pettifer, and R. Stevens. Computing identity co-reference across drug discovery datasets. In *SWAT4LS*, 2013.

[7] A. Callahan, J. Cruz-Toledo, P. Ansell, and M. Dumontier. Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. In *ESWC*. 2013.

[8] Adrien Coulet, Malika Smail-Tabbone, Amedeo Napoli, and Marie-Dominique Devignes. Ontology-based knowledge discovery in pharmacogenomics. *Adv Exp Med Biol*, 2, 2011.

[9] 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.

[10] Kevin Dalleau. biojp2rdf – a tool to rdfize biodb.jp data, under MIT licence [visited Sept. 9, 2015]: `https://github.com/KevinDalleau/biojp2rdf`.

[11] Gerben Klaas Dirk de Vries. A fast approximation of the weisfeiler-lehman graph kernel for RDF data. In *ECML-PKDD*, pages 606–621, 2013.

[12] Michel Dumontier and Natalia Villanueva-Rosales. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics*, 10(2):153–163, 2009.
[13] EBI. The EBI RDF Platform [visited Sept. 9, 2015]: `http://www.ebi.ac.uk/rdf/`.
[14] Christopher S. Funk, Lawrence E. Hunter, and K. Bretonnel Cohen. Combining heterogenous data for prediction of disease related and pharmacogenes. In *PSB*, 2014.
[15] Yael Garten, Adrien Coulet, and Russ B Altman. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*, 11(10), 2010.
[16] Yael Garten, Nicholas P. Tatonetti, and Russ B. Altman. Improving the prediction of pharmacogenes using text-derived gene-drug relationships. In *PSB*, pages 305–314, 2010.
[17] Benjamin M. Good and Mark D. Wilkinson. The Life Sciences Semantic Web is Full of Creeps! *Briefings in Bioinformatics*, 7(3):275–286, 2006.
[18] Malachi Griffith *et al.* DGIdb - mining the druggable genome. *Nature Methods*, 10:1209–10, 2013.
[19] N.T. Hansen, S. Brunak, and R.B. Altman. Generating genome-scale candidate gene lists for pharmacogenomics. *Clinical Pharmacology & Therapeutics*, pages 183–9, 2009.
[20] Philipp Heim, Steffen Lohmann, and Timo Stegemann. Interactive relationship discovery via the semantic web. In *ESWC*, volume 6088, pages 303–317, 2010.
[21] R. Hoehndorf, M. Dumontier, and G. V. Gkoutos. Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics*, 28(16), 2012.
[22] Yi Huang, Volker Tresp, Markus Bundschus, Achim Rettinger, and Hans-Peter Kriegel. Multivariate prediction for learning on the semantic web. In *ILP*, pages 92–104, 2010.
[23] Tadashi Imanishi *et al.* Hyperlink management system and ID converter system: enabling maintenance-free hyperlinks among major biological databases. *NAR*, 37:17–22, 2009.
[24] John P.A. Ioannidis. To replicate or not to replicate: The case of pharmacogenetic studies. *Circulation: Cardiovascular Genetics*, 6:413–8, 2013.
[25] Akira R. Kinjo *et al.* Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Research*, 2012.
[26] Christian Leistner, Amir Saffari, and Horst Bischof. MIForests: Multiple-instance learning with randomized trees. In *ECCV*, pages 29–42, 2010.
[27] Uta Lösch, Stephan Bloehdorn, and Achim Rettinger. Graph kernels for RDF data. In *ESWC*, pages 134–148, 2012.
[28] M. Scott Marshall *et al.* Emerging practices for mapping and linking life sciences data using rdf - a case series. *Journal of Web Semantics*, 14:2–13, 2012.
[29] Bethany Percha, Yael Garten, and Russ B. Altman. Discovery and explanation of drug-drug interactions via text mining. In *PSB*, pages 410–421, 2012.
[30] PharmGKB. Levels of evidence of annotations [visited Sept. 2, 2015]: `https://www.pharmgkb.org/page/clinAnnLevels`.
[31] Matthias Samwald *et al.* Linked open drug data for pharmaceutical research and development. *Journal of Chemoinformatics*, 3:19, 2011.
[32] Matthias Samwald *et al.* Semantically enabling pharmacogenomic data for the realization of personalized medicine. *Pharmacogenomics*, 13(2):201–12, 2012.
[33] Andreas Thor *et al.* Link prediction for annotation graphs using graph summarization. In *ISWC*, pages 714–729, 2011.
[34] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, pages 650–665, 2009.
[35] Michelle Whirl-Carrillo *et al.* Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–17, 2012.
[36] Hong-Guang Xie and Felix W Frueh. Pharmacogenomics steps toward personalized medicine. *Personalized Medicine*, 2(4):325–37, 2005.
[37] Issam Zineh, Michael Pacanowski, and Janet Woodcock. Pharmacogenetics and coumarin dosing? Recalibrating expectations. *New England Journal of Medicine*, 369:2273–5, 2013.