Integration of knowledge for personalized medicine: a pharmacogenomics case-study

Robert Hoehndorf[1], Michel Dumontier[2] and Georgios V. Gkoutos[1,3]

[1]*University of Cambridge,* [2]*Carleton University,* [3]*University of Aberystwyth*

Correspondence: rh497@cam.ac.uk, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

Abstract

The semantic integration of pharmacogenomic knowledge that is currently distributed in several resources extends the access to information that is crucial for new scientific analyses that bridge multiple resources and domains. We demonstrate how to combine formal ontological analysis and Semantic Web technology to integrate three major pharmacogenomics databases and link the combined resource to chemical and disease ontologies. We show how the additional background knowledge in these ontologies can be used to perform expressive queries in the domain of pharmacogenomics that specifically leverage the collection of formalized ontologies and enable access to knowledge that is not available in either database alone.

Introduction

The rapid increase of data generated from genetic analyses and functional genomics has necessitated the development of a number of pharmacogenomics related databases that provide invaluable resources for discovering information related to the impact of gene variations to drug responses and toxicity [1]. Such resources increasingly annotate their data with ontologies to integrate and share data across domains and resources. Ontologies provide a rich taxonomic structure and axioms that make some aspects of background domain knowledge explicit.

Different pharmacogenomics resources provide different aspects about drugs, genes, diseases, pathways and their relations. Integrating these aspects into a single framework has the potential to extend and improve the databases' utility for scientific analyses and allow them to play a crucial role in the personalization of disease intervention strategies. Here, we demonstrate how to integrate the pharmacogenomics knowledge base *PharmGKB*, DrugBank and the Comparative Toxicogenomics Database (CTD) using the Web Ontology Language (OWL) [2] such that it becomes possible to perform queries through automated reasoning over the combination of these three databases. The resulting integrated resource can be used to answer powerful queries spanning multiple databases and ontologies. In particular, we demonstrate how the integration of knowledge in pharmacogenomics with information about diseases provides a means to associate drugs with anatomical structures and systems in which they become active, and how the integration with chemical ontologies allows the reuse of chemical background knowledge to group drugs based on their chemical properties and to access their biological functions.

Methods

We have created formalized knowledge bases using OWL for several resources in pharmacogenomics, listed in Table 1. We then utilized property chains in order to infer additional information based on participation in pathways. Using the ELK OWL reasoner [3], we can classify the resulting integrated pharmacogenomics knowledge base in less than one minute and answer queries in under one second. To enable the use of these resources and the software in clinical information systems and the support of personalized treatment of disease, we make the resulting OWL knowledge bases and associated software freely available on `http://pharmgkb-owl.googlecode.com`.

Resource	Description
Anatomical Therapeutic Chemical Classification System (ATC)	Classification of drugs based on organ or system of action, therapeutic characteristics and chemical properties.
Medical Subject Headings (MESH) vocabulary	Controlled vocabulary used for indexing, cataloging, and searching for biomedical and health-related information and documents.
Comparative Toxicogenomics Database (CTD)	CTD contains manually curated data about chemical–gene/protein, gene/protein–disease and chemical–disease associations, as well as predictions based on a complex interaction network.
DrugBank	DrugBank contains detailed information about drugs and drug targets.
PharmGKB	PharmGKB contains information about the effects of human variation on drug responses.

Table 1: List of resources and software tools provided.

Results

Based on the formal representation we generate for the DrugBank, CTD and PharmGKB, we can perform an integration with other ontologies. First, we utilize mappings to ontologies of *diseases* and other *abnormalities* to perform an integration with ontologies of these domains. Such an integration allows us to use background knowledge contained in these ontologies for queries, to increase the expressivity of the representation and establish new connections between classes in PharmGKB. For example, we can use automated reasoning over the Human Disease Ontology to query for things that are associated with *Parasitic infectious disease* (`DOID:1398`) and will retrieve, among others, drugs associated with *Malaria*, *Scabies* or *Schistosomiasis* in either the CTD, PharmGKB or DrugBank. We retrieve 129 classes as result to this query, including the anti-malarial drugs *Chloroquine* and *Artemether*.

The second dimension of integrating pharmacogenomics knowledge is in respect to ontologies of chemicals and drugs. The PharmGKB, CTD and DrugBank provide references for the drugs they contain based on the ATC, the ChEBI ontology and MeSH. For this work, we generated OWL versions of the ATC and MeSH. Integration with the ATC, MeSH and ChEBI ontologies of chemicals enables expressive queries using the background knowledge contained in all three ontologies. For example, using the ChEBI ontology, we are able to query for diseases associated with some alcohol (`ChEBI:30879`) and obtain, amongst others, *Alcoholism* (`PA443309`) and *Bubonic plague* (`PA445338`) as a result. The disease *Alcoholism* is directly associated with *Ethanol* (*CHEBI:16236*), a subclass of *Alcohol* in ChEBI. *Bubonic plague*, on the other hand, is directly associated with the drug *Phenylephrine* in PharmGKB which is, in turn, a subclass of *Alcohol* in ChEBI.

PharmGKB data can list druggene and drugdisease associations which are obtained by applying the following rule: if a drug D is a component of a pathway, and that pathway has another drug, gene or disease X as component, then the drug is associated (through a pathway) with X. This kind of reasoning can be captured in OWL with *property chains*. A property chain allows to construct complex properties from simple properties by chaining two or more properties together. We have created a number of such property chains to close the knowledge in the integrated pharmacogenomics ontology against information in pathways. As a consequence, we are able to infer from the assertions that a drug participates in a pathway and this pathway has a gene as participant that there is a novel pathway-based association relation between the drug and the gene. Such property chains allows to extend our queries over the integrated resources.

Our approach of integrating resources within a domain through integration of the ontologies that provide meta-data for them is not limited to the domain of pharmacogenomics alone, but can serve as a model for other areas, including model organism databases, databases of protein functions, disease and phenotype databases. In each case, it is crucial to identify relevant biomedical ontologies based on which the content of the databases can be aligned and formalize the content of the databases in such a way that it becomes possible to answer the relevant queries. Such a model of knowledge integration can enable novel analyses that connect different domains, based on methods such as semantic similarity measures [4] or ontology enrichment analyses [5].

Acknowledgements

Funding for RH was provided by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502. Funding for MD was provided by a NSERC Discovery Grant. Funding for GVG was provided by the NIH (grant number R01 HG004838-02).

References

[1] Sim SC, Altman RB, Ingelman-Sundberg M. Databases in the area of pharmacogenetics. Hum Mutat. 2011;Available from: http://www.biomedsearch.com/nih/Databases-in-area-pharmacogenetics/21309040.html.

[2] Grau B, Horrocks I, Motik B, Parsia B, Patelschneider P, Sattler U. OWL 2: The next step for OWL. Web Semantics: Science, Services and Agents on the World Wide Web. 2008 November;6(4):309–322. Available from: http://dx.doi.org/10.1016/j.websem.2008.05.001.

[3] Kazakov Y, Krötzsch M, Simančík F. Unchain My $\mathscr{E}\mathscr{L}$ Reasoner. In: Proceedings of the 23rd International Workshop on Description Logics (DL'10). CEUR Workshop Proceedings. CEUR-WS.org; 2011. .

[4] Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Research. 2011;39(18):e119. Available from: http://nar.oxfordjournals.org/content/39/18/e119.

[5] LePendu P, Musen M, Shah N. Enabling Enrichment Analysis with the Human Disease Ontology. Journal of Biomedical Informatics. 2011;In press.