Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics

Robert Hoehndorf1, Michel Dumontier2 and Georgios V. Gkoutos1,3

1Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK
2Department of Biology, Institute of Biochemistry and School of Computer Science, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada
3Department of Computer Science, University of Aberystwyth, Old College, King Street, SY23 2AX, UK

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: Many complex diseases are the result of abnormal pathway functions instead of single abnormalities. Disease diagnosis and intervention strategies must target these pathways while minimizing the interference with normal physiological processes. Large scale identification of disease pathways and chemicals that may be used to perturb them requires the integration of information about drugs, genes, diseases and pathways. This information is currently distributed over several pharmacogenomics databases. An integrated analysis of the information in these databases can reveal disease pathways and facilitate novel biomedical analyses.

Results: We demonstrate how to integrate pharmacogenomics databases through integration of the biomedical ontologies that are used as meta-data in these databases. The additional background knowledge in these ontologies can then be used to enable novel analyses. We identify disease pathways using a novel multi-ontology enrichment analysis over the Human Disease Ontology, and we identify significant associations between chemicals and pathways using an enrichment analysis over a chemical ontology. The drug–pathway and disease–pathway associations are a valuable resource for research in disease and drug mechanisms and can be used to improve computational drug repurposing.

Contact: rh497@cam.ac.uk

1 INTRODUCTION

Pharmacogenomics aims to increase our understanding of the effect of genetic variation on the response to drugs, thereby leading to better health care through a more personalized and precise approach to medical treatment of disease. To achieve this goal, pharmacogenomics must combine and integrate data from multiple domains, including information about drug actions, gene functions, gene and protein interactions, pathways, gene expression, phenotypes, disease and genetic variation. When this information is combined, novel integrative analyses become possible that can improve our understanding of drug actions and disease mechanisms.

In complex diseases, it is often not possible to identify single aberrations underlying the disease. In order to provide possible diagnosis and treatments of such diseases, it is important that we are able to identify aberrations in the biological pathways related to such diseases in order to gain a better understanding of the synergistic molecular functions of the involved gene networks and their role in the disease. Network-based approaches can reveal specific aberrations in the processes that make up such biological systems. In particular, aberrant pathways can provide insights into the systemic imbalance underlying a disease, and can further provide targets for disease intervention (Wang et al., 2012; Chen et al., 2012). To identify aberrant pathways, access to information about pathway participants as well as their potential interactions with chemicals and diseases becomes important. One obstacle towards such an approach lies with the distribution of information across multiple heterogeneous resources.

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 (Sim et al., 2011). One prime example of such a resource is the Pharmacogenomics Knowledge Base (PharmGKB), a database in which associations between drugs, genes and their variants, and diseases is curated against primary scientific literature for which there is indication of their pharmacokinetic and pharmacodynamic properties (Thorn et al., 2010). PharmGKB’s data has been used, among many things, to extract biomedical relations from text (Coulet et al., 2011) and to make pharmacogenomic predictions such as warfarin dosing (International Warfarin Pharmacogenetics Consortium et al., 2009). Further databases that incorporate relevant pharmacogenomic knowledge include DrugBank (Knox et al., 2010), a richly annotated database of drugs and drug targets, and the Comparative Toxicogenomics Database (CTD) (Davis et al., 2010), which contains manually curated relations between chemicals, genes and diseases, and integrates them in a chemical–gene–disease network to predict novel relations.

With the advent of the Gene Ontology (GO) (Ashburner et al., 2000), ontologies are now being widely used for the annotation of data in biomedical databases, including pharmacogenomics, drug and disease databases such as the PharmaGKB, DrugBank and CTD. Ontologies aid knowledge integration by providing a rich taxonomic structure and axioms which makes some aspects of background domain knowledge explicit. Based on the generalization...
hierarchy available in ontologies, different resources can be inte-
grated even when exact matches between entities in different databases
cannot be made. Furthermore, formalized ontologies make some
aspects of a term’s meaning explicit and therefore offer the tanta-
lizing possibility to standardize biomedical knowledge and exploit
term meanings for deductive inferences that can reveal relations
across domains and levels of granularity.

Many ontology-based approaches have focused on single biome-
dical databases or domains, and they demonstrate querying, retrieval
and consistency verification within this database or domain. How-
ever, in pharmacogenomics, different databases provide different
aspects about drugs, genes, diseases, pathways and their relations,
and integrating these aspects into a single framework has the poten-
tial to extend and improve the databases’ utility for scientific analy-
ses. For example, while PharmGKB focuses on interactions between
particular gene variants and drugs, DrugBank provides compre-
ensive information about drug–gene interactions. The CTD, on
the other hand, can add information about drug–disease and gene–
disease interactions, based not only on direct evidence from primary
research literature, but also based on network-based inference of
novel association relations.

Here, we demonstrate the identification of aberrant pathways by
integrating PharmGKB with DrugBank and CTD. The integrated
pharmacogenomics knowledge base can be used to answer power-
ful queries spanning multiple ontologies and therefore transcends
domains of knowledge and levels of granularity. We demonstrate
how the link to disease ontologies enables queries for disease-
associations, and how the link with chemical ontologies allows the
reuse of chemical background knowledge to group drugs based on
their chemical properties and to access their biological functions.
We utilize these links to perform a statistical enrichment analy-
sis that reveals associations between pathways and the diseases in
which they are disturbed as well as between pathways and chemical
substances that can perturb the pathway. The integrated know-
ledge, its associated resources, the generated pathway–disease and
pathway–chemical associations, and the source code we produced in
our analysis are freely available at http://pharmgkb-owl.
googlecode.com.

MATERIALS AND METHODS

Software and ontology versions

We have incorporated several ontologies in our work. The Human Disease
Ontology (DO) is a community driven, freely available ontology that aims
to assist the integration of biomedical data that is associated with human
diseases (Chisholm et al., 2011). DO contains links to various external
terminologies such as SNOMED-CT, UMLS, ICD-9, ICD-10 and MeSH.
DO was downloaded on 19 July 2011 and contains 6,433 classes of which
21% have textual definitions. DO contains 591 classes which are fully de-
defined using external ontologies such as PATO (Gkoutos et al., 2005), the
Celltype Ontology (Bard et al., 2005), the Foundational Model of Anato-
momy (FMA) (Rosse and Mejino, 2003) and the Human Phenotype Ontology
(HPO) (Robinson et al., 2008). We further utilize the mappings to the UMLS
(Bodenreider, 2004) that are contained both in the PharmGKB and in the
Human Disease Ontology (DO) (Osborne et al., 2009b) to identify those
diseases in PharmGKB that can be directly mapped to a disease class in DO.

The ontology of Chemical Entities of Biological Interest (ChEBI) is a free
dictionary of molecular entities focused on small chemical compounds (Dog-
tyrenko et al., 2007). CHEBI classes carry a variety of information including
SMILES (Simplified Molecular Input Line Entry System) strings, IUPAC
names, and references to the Chemical Abstracts Service (CAS) Registry
Number. The ChEBI ontology was downloaded on 19 July 2011 and contains
23,589 classes, 59% of which have textual definitions.

The Anatomical Therapeutic Chemical (ATC) Classification System (Mil-
ler and Britt, 1995), controlled by the WHO Collaborating Centre for Drug
Statistics Methodology (WHOCC), provides a drug classification based on a
grouping according to the organ or system they act upon as well as their the-
rapeutic and chemical characteristics. ATC was downloaded from the KEGG
website on 17 July 2011 and contains 10,167 classes.

PharmGKB data were downloaded from the PharmGKB website on 1
October 2011 and contain information about 3,004 drugs, 3,203 diseases,
27,421 genes and their relationships as well as 1,408 pathways (including
pathways obtained from resources such as the Pathway Interaction Data-
base). The CTD is a database containing relations between chemical entities,
genesis and diseases, and we downloaded the files on 1 October 2011. The
DrugBank database was downloaded on 1 October 2011.

DrugBank and CTD use the MeSH thesaurus to provide identifiers for
drugs and diseases. In the absence of a publicly available OWL version of
MeSH, we wrote a software to generate such an OWL representation and
make this software as well as the resulting OWL ontology freely availa-
able (http://pharmgkb-owl.googlecode.com). The OWL version of
MeSH that we generate represents MeSH’s taxonomic identifiers as
OWL classes and taxonomic relations between them as subclass relations.
Although different formalizations have been discussed that use, for exam-
ple, the SKOS vocabulary and broader-than and narrower-than relations,
the representation we chose is motivated by our use case and evaluation
procedure.

A large number of tools are available to perform enrichment analyses
(Subramanian et al., 2005) over ontologies. The FUNC tool (Prufer et al.,
2007) is a generic ontology enrichment tool in that it is able to perform
different kinds of tests (including correction for multiple testing) and
supports the use of arbitrary graph structures to perform enrichment analy-
ses. The graph structures must be represented in a format that corresponds
to the database format of GO. To enable FUNC to use graph structures of
ontologies that are being developed using OWL, we implemented the
OntoFUNC software which generates these graph structure representations
based on OWL ontologies. The OntoFUNC tool is based on FUNC version
Currently, OntoFUNC utilizes only the subclass relations in an OWL ontol-
ogy to generate a graph representation of an ontology’s taxonomy. To
generate an ontology’s taxonomy, we utilize the ELK reasoner (Kaz-
akov et al., 2011) and perform queries for subclasses of each named class
in the OWL ontology. We make OntoFUNC freely available on http:
//ontofunc.googlecode.com.

We have used the Groovy language to implement OntoFUNC and the soft-
ware to integrate the pharmacogenomics databases. Our software makes use
of the OWLAPI (Horridge et al., 2007) to generate OWL 2 ontologies from
the PharmGKB. The source code to perform the OWL-based integration and
analysis is freely available on our project website. Furthermore, we intend to
update the resources and analysis results at least four times a year.

Upper ontology

We use OWL to formalize parts of the PharmGKB database, DrugBank
and CTD. Since we integrate large amounts of information, we restrict our repre-
sentation to the OWL EL profile (Motik et al., 2009) which enables tractable
automated reasoning (Hochdorf et al., 2011a), but does not support, among
other OWL constructs, the use of inverse, functional or symmetric object
properties. We then demonstrate the integration of the resulting formaliza-
tion with the Human Disease Ontology (DO) (Osborne et al., 2009a), the
ChEBI ontology, the Anatomical Therapeutic Chemical (ATC) classification
and the Modern Subject Headings (MeSH) thesaurus. Based on the result-
ing integrated ontology, we then show how property chains can be used to
enrich the knowledge available in these databases, demonstrate powerful
queries, describe our tool OntoFUNC and apply OntoFUNC for ontology enrichment over DO in order to evaluate the representation we develop.

From the information that is present in PharmGKB, CTD and DrugBank, we focus on the information about genes, drugs, diseases and pathways in our formalization, and we assert relations between them using complex OWL axioms.

Genes in PharmGKB represent both genes and their products, i.e., no explicit distinction is made between them. Therefore, PharmGKB's gene entries are linked both to reference databases for genes (Entrez) as well as gene products (UniProt). Some genes are linked to specific gene variants (through a RefSNP identifier). In the PharmGKB, drugs are chemical entities and may contain links to external databases such as the ChEBI ontology, ATC or related pharmacogenomics resources such as DrugBank. Diseases in PharmGKB are linked to external disease terminologies, including SNOMED-CT, UMLS (Bodenreider, 2004), ICD-9, ICD-10 and DO. Pathways in PharmGKB are sets of interactions that occur between genes, proteins and drugs based on interdependent relationships and events between the pathways' participants. In PharmGKB, some pathways are further associated with diseases. We use the classes from PharmGKB as the foundation of our work: Drug, Disease, Gene and Pathway. The four upper-level classes are declared as disjoint, and we align the classes found in DrugBank and CTD on these four classes.

Based on the distinctions made explicit in the PharmGKB, DrugBank and CTD, and we automatically generate a new class for each entry in these databases and create an axiom that makes this class a subclass of either Gene, Drug, Disease or Pathway. For example, the identifier PA28242 is in PharmGKB represents the class of FOXP2 genes in PharmGKB. We create a class PA28242 as a subclass of Gene and we label that class FOXP2. The identifier PA162263534 in PharmGKB, on the other hand, represents the class of Ototoxicity, and we create the class PA162263534 as a subclass of Disease and label this class Ototoxicity. CTD uses MeSH identifiers to refer to drugs and diseases, and we do not add new classes but directly use the classes in our OWL representation of MeSH.

**Relations**

Relations can be established between any of the classes in PharmGKB, CTD and DrugBank based on the interactions ascribed by the database curators. These relations are derived from statements in the literature and include relations between genes and their variants, drugs, diseases and pathways (Thorn et al., 2010). The literature descriptions of these relationships capture a great variety of biologically diverse interactions. Such relationships can be further characterised based on the type of entities they refer to (including their domain and range) and the type of interaction between them. For example, specific relations could be created to denote gene-to-gene interactions or gene-to-disease interactions, while other relation types may characterize the type of interactions between them.

Currently, specific relationships are not available in all the databases. PharmGKB provides association relations, but not the mode of interaction between drugs and genes. Consequently, we have introduced the relation **directly-associated-with** and used this to formalize all the relationships that are directly asserted in either of the integrated databases. In particular, when we find an asserted association between a pair of entities (X, Y) in PharmGKB, DrugBank or CTD, we assert the OWL axiom X SubClassOf: directly-associated-with some Y. The relation **directly-associated-with** is a sub-relation of **associated-with**, which includes other types of association relations (in particular, associations that are a consequence of participation in a pathway).

A second source of information about relations between drugs, genes and diseases is available in PharmGKB's representation of pathways. PharmGKB contains pathway descriptions both from external resources and manually curated pharmacologically relevant pathways. In these pathways, genes, drugs and diseases may be **participants**. Further details about pathways are available, including details about specific events and reactions, such as the inputs and outputs of reactions, or phenotypes resulting from events. Currently, we formalize only pathway components (i.e., its participants) and leave the formalization of reactions and events as future work.

We introduce six relations of the types **has-participant** and **participates-in** to distinguish different types of participation in a pathway: drug–pathway, gene–pathway, disease–pathway, pathway–drug, pathway–gene and pathway–disease relations. Each relation is further restricted by domain and range assertions. For example, the drug–pathway relation (labelled **drug-participates-in-pathway**) has its domain restricted to **Drug** and its range to **Pathway**.

Based on these relations, we automatically create a formal description of the pathways found in PharmGKB. For example, if we find a pathway **P** described in PharmGKB that includes a drug **D** as component, we create the two OWL axioms:

D SubClassOf: drug-participates-in-pathway some P
P SubClassOf: pathway-has-participant-drug some D

These axioms enable us to retrieve the pathway participants and further distinguish their types. Although all pathways are also available in the BioPAX format (Demir et al., 2010), which is an OWL-based representation format, we do not use the BioPAX representation since it uses complex OWL constructs that go beyond OWL EL.

**Mapping of PharmGKB to DO**

In PharmGKB, we identify 1,823 diseases that can be mapped to DO. To enable automated reasoners to utilize the mappings between PharmGKB's disease classes and DO, we include DO in the knowledge base we create and generate an equivalent class axiom for each mapping. For example, Hodgkin disease (PA444485) is linked to the UMLS concept identifier C0019829, and the DO class Hodgkin's lymphoma (DOID:8567) is linked to the same UMLS concept identifier. Based on this link, we generate the equivalent classes axiom PA444485 EquivalentTo: DOID:8567. We further create equivalent class axioms between MeSH's disease classes and DO classes when DO provides the information as a cross-reference.

**RESULTS**

**Integration with disease ontologies**

Based on the representation we generated 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 DrugBank, CTD and PharmGKB. Several formal disease representations are available and are potential candidates for integrating pharmacogenomics knowledge with representations of diseases. We base our work on DO, since it is freely available and actively maintained.

The mappings from PharmGKB and CTD to DO allow us to use the additional knowledge contained in DO for querying the content of the databases. For example, we can use inference over DO to query for disease classes that are not available in the PharmGKB: a query for things that are associated with Parasitic infectious disease (DOID:1398) will retrieve, among others, drugs associated with Malaria, Scabies or Schistosomiasis. We retrieve 129 classes as result to our query of things that are associated with Parasitic infectious diseases, including the anti-malarial drugs Chloroquine and Artemether.
As a further extension of the query capabilities, we can make use of the class definitions that were developed for DO. These definitions can link diseases to the agent that causes the disease, or to the disease’s anatomical location. In DO, the infectious agents are specified through the NCBI Taxonomy, while the anatomical location is characterized through the ontology of Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003). For example, the class *Chikungunya* (DOID:0050012) is defined as a viral infectious disease caused by a *Chikungunya virus* (NCBITaxon:37124) and located in the Joint (FMA:7490). Using these definitions, we can ask for drugs that are associated with joint diseases using the query class Drug and directly-associated-with some (Disease and located-in some Joint). As a result to this query, we obtain drugs such as *Folic acid* (PA449692) that are used to treat Arthritis.

The link to DO does not only enable powerful, new queries over the information contained in pharmacogenomics databases, but further enables the possibility for enrichment analyses using DO (LePendu et al., 2011). The additional links that are established between DO and organism taxonomy as well as between DO and anatomy ontologies can be used to further refine enrichment analyses. Furthermore, these links enable enrichment for classes from the linked ontologies. For example, using DO’s links to the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003), drugs can be grouped based on the organs or tissue they interact with or grouped based on anatomical systems in which they are active.

**Integration with chemical ontologies**

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. The ChEBI ontology is available in the OBO Flatfile Format (Horrocks, 2007) and can be integrated into OWL ontologies. ATC, however, is not publicly available in a format that is compatible with OWL. Therefore, we generated an OWL-based representation of ATC ourselves and integrate this representation in the knowledge base we create. Similar to the mappings between diseases in PharmGKB and CTD, we generate equivalent class axioms when the PharmGKB or DrugBank link a drug to a class in ChEBI. For example, the drug *Mercaptopurine* (PA450379 in PharmGKB and DB01033 in DrugBank) is linked to the ChEBI class *Purine-6-thiol* (CHEBI:22208), and based on this information we create an equivalent class axiom between the class PA450379 and CHEBI:22208 as well as between DB01033 and CHEBI:22208. We further use the cross-references provided by the CTD to create equivalent class axioms between chemicals in MeSH and CHEBI.

Both the ATC and the PharmGKB distinguish between drug classes and small molecules/drugs. In many cases, the link between a specific drug or small molecule in the PharmGKB and the ATC does not reflect an assertion of equivalence, but rather a subclass assertion. For example, *Mercaptopurine* is linked to *Purine analogues (ATC:L01BB)* in the ATC. On the other hand, links between drug classes in the PharmGKB and drug classes in ATC usually represent assertions of equivalence. For example, the class *Purine analogues* (PA452634) in PharmGKB is linked to the corresponding class in ATC with the intention that both are equivalent classes. Based on these observations, we treat the link between PharmGKB and ATC drug classes as assertions of equivalence, and the link between drugs/small molecules and an ATC class as a subclass assertion.

Integration with the ATC and ChEBI ontologies of chemicals enables expressive queries using the background knowledge contained in both 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 (PA450935) which is mapped to ChEBI:8093 which is, in turn, a subclass of Alcohol in ChEBI.

We can further use the relations that are asserted in the ChEBI ontology to further retrieve specific classes. For example, we can ask for diseases that are associated with mutagenic drugs acting on the central nervous system using the classes *Mutagen* (ChEBI:25435) and *Central nervous system drug* (ChEBI:35470) to ask for subclass of:

Disease and directly-associated-with some (has-role some Mutagen and has-role some 'Central nervous system drug')

The results of this query include a range of diseases and associated genes that are found in CTD and PharmGKB and are directly associated with the mutagenic central nervous system drug Caffeine. For example, CTD results would include Liver Cirrhosis (MESH:D008106) and Anxiety disorders (MESH:D001008) that are linked to the ADORA2A gene, whilst the PharmGKB results would include Schizophrenia (PA447216). By examining the manually curated drug-gene interactions from DrugBank, we retrieve a variety of genes that are linked to Caffeine such as the PDE4B gene (Entrez Gene 5142), that neither PharmGKB nor CTD include in their known drug-gene interactions. While PharmGKB links Schizophrenia to Caffeine, the additional information available from DrugBank reveals the mechanism and gene based on which the disease and drug are associated: in a recent study, the gene encoding PDE4B was reported to be disrupted in a subject diagnosed with schizophrenia and a relative with chronic psychiatric illness (Millar et al., 2008), and together with the interaction between PDE4B and Caffeine, this gene and its interactions provide the evidence for the link between Caffeine and Schizophrenia. This connection cannot be discovered by examining either of the three resources independently, and finding evidence for this connection has the potential to provide a biological explanation of the relatively unknown role that caffeine plays in patients that suffer from Schizophrenia (Martin et al., 2008).

**Integration with pathway knowledge**

We can further extend the set of drug–gene and drug–disease associations that are available by applying the following rule: if a drug *D* is a component of a pathway *P*, and that pathway has another drug, gene or disease *X* as component, then the drug is associated with *X* (via the pathway *P*). 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. For example, when we want to infer from the assertions that, if a drug participates in a
that are only significant as a result of their subclasses’ being significant. To identify chemicals enriched in pathways, we use a similar method based on an enrichment analysis over the ChEBI ontology and using the query

\[
\text{Drug and associated-with some } G
\]
to identify chemicals that are associated with genes.

Using a \( p \)-value of 0.05 as measure of a significant association, we identify 22,653 significant pathway–disease associations, out of which 6,304 are over-represented disease classes in a pathway and 16,349 are under-represented disease classes. We further identify 13,826 significant pathway–chemical associations, out of which 12,564 are over-represented chemical classes for a pathway and 1,262 are under-represented chemical classes.

As one example, we explored the list of disease pathways and found that the participants of the Zidovudine Pathway (PharmGKB:PA165859361) are strongly over-represented for Mood disorder (DOID:3324) and the central compound of the pathway Zidovudine. Zidovudine is a nucleoside reverse transcriptase inhibitor (NRTI) administered to patients suffering from serious manifestations of HIV infections with acquired immunodeficiency syndrome (AIDS) or AIDS-related complex (ARC) (Arts et al., 1998; Lewis et al., 2001). Known side effects of Zidovudine include fatigue, headache, and myalgia as well as malaise and anorexia which clearly demonstrate the association of Zidovudine with Mood disorders (Frissen et al., 1994; Max and Sherer, 2000).

As a second example, we investigated the pathway-chemical associations and identified a strong association of the drug clopidogrel (CHEBI:37941) with an Endothelin signaling pathway (PharmGKB:PA164728163). Clopidogrel is a thienopyridine-derived anti-platelet drug that inhibits platelet aggregation and prolongs bleeding time (Savi et al., 1994; Herbert et al., 1993). It is administered for inhibiting blood clots thereby preventing ischemic events such as cardiovascular death, myocardial infarction, or stroke in atherothrombotic patients (Bhatt and Topol, 2003). Clopidogrel results in inhibition of platelet activation due to clopidogrel’s antagonism effect on the platelets’ adenosine diphosphate (ADP) receptors (Yang and Fareed, 1997). Furthermore, clopidogrel has been shown to inhibit smooth muscle cell mitogenesis (Bhatt and Topol, 2003) and modulate vascular smooth muscle (Yang and Fareed, 1997). In particular, it has been shown in rats and rabbits that it inhibits the serotonin and endothelin-1 mediated vascular smooth muscle contraction (Yang and Fareed, 1997) thereby perturbing the endothelin signaling pathway (PharmGKB:PA164728163).

The full dataset, including the significant associations, the plain and the corrected \( p \)-values for all associations are provided at the project website (http://ontofunc.googlecode.com). To enable further analysis, we also make the source code freely available.

Implementation and availability

We integrated PharmGKB, DrugBank and CTD databases using a set of patterns based on which a large part of the content of PharmGKB, DrugBank and CTD can be expressed in OWL. To enable the integration of these databases, we generated an OWL version of the MeSH thesaurus and the ATC, and we further utilized the ChEBI ontology and DO in the integration. Application of our software yields an ontology that contains more than 650,000
classes, 93 object properties, more than 3,200,000 subclass axioms and more than 75,000 equivalent classes axioms. The software that implements these patterns and converts the databases, the software to generate OWL versions of MeSH and ATC, and the resulting ontology files are freely available on our project’s website at http://pharmgkb-owl.googlecode.com. Table 1 lists the resources and software tools we provide in order to integrate and analyze knowledge in pharmacogenomics.

The ontology we create falls in the OWL EL fragment of OWL (Motik et al., 2009) and consequently allows for tractable automated reasoning using reasoners which are optimized for OWL EL (Hoehndorf et al., 2011a; Kazakov et al., 2011). Using the ELK reasoner (Kazakov et al., 2011), the ontology we create classifies in less than one minute on hardware consisting of two Intel® Xeon® 2.4GHz quad-core CPUs with 24GB memory.

**DISCUSSION**

Relevant knowledge about pharmacogenomics is distributed across several databases, each of which focuses on different aspects of this complex domain. For example, in the PharmGKB, no extensive information about drug–disease or gene–disease associations is provided, but this information is contained in other databases such as DrugBank and CTD.

Here, we used the PharmGKB, CTD and DrugBank to demonstrate how the combination of Semantic Web technologies and formal ontological analysis can be used to integrate different resources relevant for pharmacogenomics research. This integration does not only lead to significantly improved capabilities for knowledge retrieval, but also enables statistical analyses of the data in these databases that reveal associations between pathways and diseases as well as associations between pathways and chemicals. While our approach is currently limited by the depth and reliability of the data contained in the integrated databases, both as a consequence of a lack of complete knowledge on biological pathways and a lack of complete curation of literature, methods of knowledge integration can provide the means to generate insights that lead to more direct biological investigation.

Biomedical ontologies and Semantic Web technology play a crucial role in such integration methods. Ontologies provide a rich taxonomic structure and axioms that can provide background knowledge based on which knowledge from different domains can be integrated. In particular, ontologies allow for a generalization of concepts that is useful when exact matches between entities in different resources are not possible and when a grouping of classes based on various features is required. For example, through the link to DO and the classification it provides, we are able to identify specific neoplasms using the additional information provided by DO’s classification of diseases. Additionally, relations in ontologies can be combined with information from the pharmacogenomics databases in order to compose complex associations that cannot be found within either database alone. For example, the background knowledge in DO allows us to group drugs based on the anatomical site at which they are active. This information may lead to additional insights into the drugs’ mechanisms of action.

In order to utilize this background knowledge, it is necessary to automatically process biomedical ontologies and extract relevant knowledge. Semantic Web technology, in particular automated reasoning software, is now capable of processing large biomedical ontologies, to process, verify and query OWL knowledge bases. Recent progress in efficient automated reasoning, in particular related to reasoning over large life science ontologies in the OWL EL profile (Hoehndorf et al., 2011a; Kazakov et al., 2011) has enabled the potential to utilize automated reasoning in software application and scientific data analyses, and our performance results of automated reasoning over the integrated pharmacogenomics databases

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>OWL version available from</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical Therapeutic Chemical Classification System (ATC)</td>
<td>Classification of drugs based on organ or system of action, therapeutic characteristics and chemical properties.</td>
<td><a href="http://pharmgkb-owl.googlecode.com">http://pharmgkb-owl.googlecode.com</a></td>
</tr>
<tr>
<td>Medical Subject Headings (MESH vocabulary)</td>
<td>Controlled vocabulary used for indexing, cataloging, and searching for biomedical and health-related information and documents.</td>
<td><a href="http://pharmgkb-owl.googlecode.com">http://pharmgkb-owl.googlecode.com</a></td>
</tr>
<tr>
<td>OntoFUNC</td>
<td>Enables enrichment analyses over arbitrary OWL ontologies using the FUNC tool.</td>
<td><a href="http://ontofunc.googlecode.com">http://ontofunc.googlecode.com</a></td>
</tr>
</tbody>
</table>
demonstrate that it is even feasible to implement real-time query and analysis applications based on automated reasoning.

Our approach is not limited to the domain of pharmacogenomics alone, but can serve as a model for integrative analyses in other areas, including model organism databases (Hoehndorf et al., 2011c), databases of protein functions, disease, phenotype databases (Hoehndorf et al., 2011d) or databases of computational models (Hoehndorf et al., 2011b). In each case, it is crucial to identify relevant biomedical ontologies based on which the content of the databases can be aligned, formalize the content of the databases in such a way that it becomes possible to answer the relevant queries and perform the desired analysis. Such a model of knowledge integration may enable novel analyses that connect different domains, based on methods such as semantic similarity measures or ontology enrichment analyses, and thereby provide a general approach towards integrative bioinformatics analyses.

ACKNOWLEDGMENTS

The OntoFUNC tool was developed at the BioHackathon 2011, which took place in Kyoto, Japan. Funding for RH was provided by the European Commission’s 7th Framework Programme, RICORDO project, grant number 248502. Funding for GVG was provided the NIH (grant number R01 HG004838-02). Funding for MD was provided by an NSERC Discovery Grant.

REFERENCES


Gkiontas, G. V. et al. (2005). Using ontologies to describe mouse phenotypes. Genome Biology, 6(1).


Hoehndorf, R. et al. (2011c). Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. PLOS ONE, 6(7), e22006.


