BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 16 2012, pages 2169–2175
doi:10.1093/bioinformatics/bts350

Databases and ontologies Advance Access publication June 17, 2012

Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics

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

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

Associate Editor: 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.

Availability: http://pharmgkb-owl.googlecode.com

Contact: rh497@cam.ac.uk

Received on February 8, 2012; revised on June 11, 2012; accepted on June 12, 2012

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 (Chen *et al.*, 2012; Wang *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 have 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 integrated even when exact matches between entities in different databases cannot be

*To whom correspondence should be addressed.

made. Furthermore, formalized ontologies make some aspects of a term's meaning explicit and therefore offer the tantalizing 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 biomedical databases or domains, and they demonstrate querying, retrieval and consistency verification within this database or domain. However, 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 potential to extend and improve the databases' utility for scientific analyses. For example, while PharmGKB focuses on interactions between particular gene variants and drugs, DrugBank provides comprehensive 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 on network-based inference of novel association relations.

Herein, we demonstrate the identification of aberrant pathways by integrating PharmGKB with DrugBank and CTD. The integrated pharmacogenomics knowledge base can be used to answer powerful 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 use these links to perform a statistical enrichment analysis 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 knowledge, 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.

2 MATERIALS AND METHODS

2.1 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 are associated with human diseases (Chrisholm *et al.*, 2011). DO contains links to various external terminologies such as SNOMED-CT, UMLS, ICD-9, ICD-10 and Medical Subject Headings (MeSH). DO was downloaded on July 19, 2011 and contains 6433 classes of which 21% have textual definitions. DO contains 591 classes that are fully defined using external ontologies such as PATO (Gkoutos *et al.*, 2005), the Celltype Ontology (Bard *et al.*, 2005), the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) and the Human Phenotype Ontology (Robinson *et al.*, 2008). We further use the mappings to the UMLS (Bodenreider, 2004) that are contained both in the PharmGKB and in the 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 (Degtyarenko *et al.*, 2007). ChEBI classes carry a variety of information including (Simplified Molecular Input Line Entry System SMILES) strings, IUPAC names and references to the Chemical Abstracts Service Registry Number. The ChEBI ontology was downloaded on July 19, 2011, and contains 23 589 classes, 59% of which have textual definitions.

The Anatomical Therapeutic Chemical (ATC) Classification System (Miller 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 therapeutic and chemical characteristics. ATC was downloaded from the KEGG web site on July 17, 2011, and contains 10 167 classes.

PharmGKB data were downloaded from the PharmGKB web site on October 1, 2011, and contain information about 3004 drugs, 3203 diseases, 27 421 genes and their relationships as well as 1408 pathways (including pathways obtained from resources such as the Pathway Interaction Database). The CTD is a database containing relations between chemical entities, genes and diseases, and we downloaded the files on October 1, 2011. The DrugBank database was downloaded on October 1, 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 available (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 example, 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 four different kinds of tests (including correction for multiple testing) and supports the use of arbitrary graph structures to perform enrichment analyses. 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 0.4.4, downloaded on August 25, 2011, from http://func.eva.mpg.de. Currently, OntoFUNC use only the subclass relations in an OWL ontology to generate a graph representation of an ontology's taxonomy. To generate an OWL ontology's taxonomy, we use the ELK reasoner (Kazakov *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 software 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 web site. Furthermore, we intend to update the resources and analysis results at least four times a year.

2.2 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 representation to the OWL EL profile (Motik *et al.*, 2009) which enables tractable automated reasoning (Hoehndorf *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 formalization with the DO (Osborne *et al.*, 2009a), the ChEBI ontology, the ATC classification and the MeSH thesaurus. Based on the resulting 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, 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* 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.

2.3 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 characterized 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.

2.4 Mapping of PharmGKB to DO

In PharmGKB, we identify 1823 diseases that can be mapped to DO. To enable automated reasoners to use 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.

3 RESULTS

3.1 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 use 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 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 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.

3.2 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:2208) and based on this information we create an equivalent class axiom between the class PA450379 and CHEBI:2208 as well as between DB01033 and CHEBI:2208. 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, among 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, while 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. Although 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.*, 2005), 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 who suffer from schizophrenia (Martin *et al.*, 2008).

3.3 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 pathway and the pathway has a gene as participant, that this drug is associated with the gene (via the pathway), we first construct the complex

relation 'drug-participates-in-pathway'-followed-by-'pathway-has-participant-gene', and assert this complex relation as a sub-relation of the new relation 'pathway-associated-with':

```
drug-participates-in-pathway o
   pathway-has-participant-gene
     -> pathway-associated-with
```

We declare the relation 'pathway-associated-with' as a sub-relation of 'associated-with' so that we are able to use 'associated-with' for queries over all types of direct association in PharmGKB, 'pathway-associated-with' for the specific case that an entity is associated with another entity through participation in a common pathway, and 'directly-associated-with' for associations that have directly been declared in either database.

Based on the property chain, we infer from the assertions that a drug participates in a pathway and this pathway has a gene as participant that there should be an association relation between the drug and the gene. For example, from the PharmGKB we obtain the information that doxorubicin participates in the doxorubicin pathway, pharmacokinetics pathway (`PA165292177`). This pathway has, as one of its participants, the gene *AKR1C3* (`PA24679`). Based on the property chain, we added to the knowledge base, we infer, through inference over participation in the pathway, that *AKR1C3* is pathway associated with the drug doxorubicin.

We can then use this property chain to extend our queries. Querying PharmGKB for things directly associated with doxorubicin gives 411 classes as a result. When the property chain is applied (using associated-with), we obtain 446 results, including the genes that participate in the doxorubicin pathway. We include property chains in our OWL representation to close the pathway–association relation with respect to participation in the same pathway.

3.4 OntoFUNC: identifying aberrant pathways

The structure of biomedical ontologies is not only a valuable feature to enable retrieval and querying but is widely used in the form of enrichment analyses to analyze, for example, gene expression (Subramanian *et al.*, 2005). An enrichment analysis uses the graph structure of an ontology, such as GO, to determine whether a defined set of genes shows statistically significant, concordant differences between two biological states; it uses the annotation of a set of genes with GO terms and the GO graph structure and inference rules to statistically test for enriched GO terms.

We developed the OntoFUNC software, an extension of the popular FUNC enrichment tool (Prufer *et al.*, 2007), and applied it to identify disease classes (from DO) and chemical classes (from ChEBI) that are enriched in pathways. We first identify, for each gene *G* contained in our combined knowledge base, the diseases associated with *G* by using a subclass query of the form:

```
Disease and associated-with some G
```

For each pathway *P*, we then identify the genes that participate in the pathway and use the hypergeometric test of the FUNC tool to identify diseases that are enriched within the set of genes in the pathway. We use FUNC's option to correct for multiple testing using a control of the family-wise error rate. Furthermore, we use FUNC's refinement operation that removes those significant classes 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

```
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 6304 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 1262 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 administered to patients suffering from serious manifestations of HIV infections with acquired immunodeficiency syndrome (AIDS) or AIDS-related complex (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 signalling pathway (`PharmGKB:PA164728163`). Clopidogrel is a thienopyridine-derived anti-platelet drug that inhibits platelet aggregation and prolongs bleeding time (Herbert *et al.*, 1993; Savi *et al.*, 1994). 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 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 signalling 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 web site (http://ontofunc.googlecode.com). To enable further analysis, we also make the source code freely available.

3.5 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 use the ChEBI ontology and DO in the integration. Application of our software yields an ontology that contains >650 000 classes, 93 object

Table 1. List of resources and software tools provided

Resource	Description	OWL version available from
Anatomical Therapeutic Chemical Classification System (ATC)	Classification of drugs based on organ or system of action, therapeutic characteristics and chemical properties.	`http://pharmgkb-owl.googlecode.com`
MESH vocabulary	Controlled vocabulary used for indexing, cataloging and searching for biomedical and health-related information and documents.	`http://pharmgkb-owl.googlecode.com`
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.	`http://pharmgkb-owl.googlecode.com`
DrugBank	DrugBank contains detailed information about drugs and drug targets.	`http://pharmgkb-owl.googlecode.com`
PharmGKB	PharmGKB contains information about the effects of human variation on drug responses.	`http://pharmgkb-owl.googlecode.com`
OntoFUNC	Enables enrichment analyses over arbitrary OWL ontologies using the FUNC tool.	`http://ontofunc.googlecode.com`

properties, >3 200 000 subclass axioms and >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 web site at http://pharmgkb-owl.googlecode.com. Table 1 lists the resources and software tools we provide in order to integrate and analyse 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 <1 min on hardware consisting of two Intel® Xeon® 2.4 GHz quad-core CPUs with 24 GB memory.

4 DISCUSSION

Relavant 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.

Herein, 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. Although 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. In additional, 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 use 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 use automated reasoning in software application and scientific data analyses, and our performance results of automated reasoning over the integrated pharmacogenomics databases 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.

ACKNOWLEDGMENT

The OntoFUNC tool was developed at the BioHackathon 2011, which took place in Kyoto, Japan.

Funding: European Commission's 7th Framework Programme (grant number 248502 to RH); National Institutes of Health (grant number R01 HG004838-02 to GVG); National Sciences and Engineering Research Council of Canada (grant number XX to MD)

Conflict of Interest: none declared.

REFERENCES

Arts,E.J. *et al.* (1998) 3′-azido-3′-deoxythymidine (azt) mediates cross-resistance to nucleoside analogs in the case of azt-resistant human immunodeficiency virus type 1 variants. *J. Virol.*, **72**, 4858–4865.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**.

Bard,J. *et al.* (2005) An ontology for cell types. *Genome Biol.*, **6**.

Bhatt,D.L. and Topol,E.J. (2003) Scientific and therapeutic advances in antiplatelet therapy. *Nat. Rev. Drug Discov.*, **2**, 15–28.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**.

Chen,L. *et al.* (2012) Integrative network analysis to identify aberrant pathway networks in ovarian cancer. *Pacific Symp. Biocomput.*, **17**, 31–42.

Chrisholm,R. *et al.* (2011) Disease ontology.

Coulet,A. *et al.* (2011) Integration and publication of heterogeneous text-mined relationships on the semantic web. *J. Biomed. Semant.*, **2**(Suppl. 2), S10+.

Davis,A.P. *et al.* (2010) The comparative toxicogenomics database: update 2011. *Nucleic Acids Res.*

Degtyarenko,K. *et al.* (2007) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*

Demir,E. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.

Frissen,P.H. *et al.* (1994) Zidovudine and interferon-alpha combination therapy versus zidovudine monotherapy in subjects with symptomatic human immunodeficiency virus type 1 infection. *J. Infect. Dis.*, **169**, 1351–1355.

Gkoutos,G.V. *et al.* (2005) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**.

Herbert,J.M. *et al.* (1993) Inhibitory effect of clopidogrel on platelet adhesion and intimal proliferation after arterial injury in rabbits. *Arterioscler. Thromb.*, **13**, 1171–1179.

Hoehndorf,R. *et al.* (2011a) A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics*, **27**, 1001–1008.

Hoehndorf,R. *et al.* (2011b) Integrating systems biology models and biomedical ontologies. *BMC Sys. Biol.*, **5**, 124+.

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

Hoehndorf,R. *et al.* (2011d) Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.*, **39**, e119.

Horridge,M. *et al.* (2007) Igniting the OWL 1.1 touch paper: The OWL API. In *Proceedings of OWLED 2007: Third International Workshop on OWL Experiences and Directions*.

Horrocks,I. (2007) OBO flat file format syntax and semantics and mapping to OWL Web Ontology Language. *Technical report*, University of Manchester. http://www.cs.man.ac.uk/horrocks/obo/.

International Warfarin Pharmacogenetics Consortium *et al.* (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med.*, **360**, 753–764.

Kazakov,Y. *et al.* (2011) Unchain my $\mathcal{EL}$ reasoner. In *Proceedings of the 23rd International Workshop on Description Logics (DL'10)*, CEUR Workshop Proceedings. CEUR-WS.org.

Knox,C. *et al.* (2010) Drugbank 3.0: a comprehensive resource for omics research on drugs. *Nucleic Acids Res.*

LePendu,P. *et al.* (2011) Enabling enrichment analysis with the human disease ontology. *J. Biomed. Inform.*. In press.

Lewis,W. *et al.* (2001) Combined antiretroviral therapy causes cardiomyopathy and elevates plasma lactate in transgenic aids mice. *Lab. Invest.*, **81**, 1527–1536.

Martin,C.A. *et al.* (2008) Caffeine use: association with nicotine use, aggression, and other psychopathology in psychiatric and pediatric outpatient adolescents. *ScientificWorldJournal*, **8**, 512–516.

Max,B. and Sherer,R. (2000) Management of the adverse effects of antiretroviral therapy and medication adherence. *Clin. Infect. Dis.*, **30**(Suppl. 2), S96–116.

Millar,J.K. *et al.* (2005) DISC1 and PDE4B are interacting genetic factors in schizophrenia that regulate cAMP signaling. *Science*, **310**, 1187–1191.

Miller,G.C. and Britt,H. (1995) A new drug classification for computer systems: the atc extension code. *Inter. J. Bio-Med. Comput.*, **40**, 121–124. (Asia Pacific Association of Medical Informatics (APAMI)).

Motik,B. *et al.* (2009) Owl 2 web ontology language: Profiles. Recommendation, World Wide Web Consortium (W3C).

Osborne,J. *et al.* (2009a) Annotating the human genome with disease ontology. *BMC Genomics*, **10**(Suppl. 1), S6+.

Osborne,J. *et al.* (2009b) Annotating the human genome with disease ontology. *BMC Genomics*, **10**(Suppl. 1), S6+.

Prufer,K. *et al.* (2007) Func: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics*, **8**, 41+.

Robinson,P.N. *et al.* (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Rosse,C. and Mejino,J.L.V. (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J. Biomed. Inform.*, **36**, 478–500.

Savi,P. *et al.* (1994) The antiaggregating activity of clopidogrel is due to a metabolic activation by the hepatic cytochrome p450-1a. *Thromb. Haemost.*, **72**, 313–317.

Sim,S.C. *et al.* (2011) Databases in the area of pharmacogenetics. *Hum. Mutat.*

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. USA*, **102**, 15545–15550.

Thorn,C.F. *et al.* (2010) Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*, **11**, 501–505.

Wang,J. *et al.* (2012) Identification of aberrant pathways and network activities from high-throughput data. *Brief. Bioinform.*

Yang,L.H. and Fareed,J. (1997) Vasomodulatory action of clopidogrel and ticlopidine. *Thromb. Res.*, **86**, 479–491.