A Semantic Web Ontology for Small Molecules and Their Biological Targets

JooYoung Choi,[†,‡] Melissa J. Davis,[†,§] Andrew F. Newman,[†] and Mark A. Ragan*[,†,‡]

Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, The University of Queensland, Brisbane, QLD 4072, Australia, and Queensland Facility for Advanced Bioinformatics, Queensland Bioscience Precinct, Brisbane, QLD 4072, Australia

Received November 26, 2009

A wide range of data on sequences, structures, pathways, and networks of genes and gene products is available for hypothesis testing and discovery in biological and biomedical research. However, data describing the physical, chemical, and biological properties of small molecules have not been well-integrated with these resources. Semantically rich representations of chemical data, combined with Semantic Web technologies, have the potential to enable the integration of small molecule and biomolecular data resources, expanding the scope and power of biomedical and pharmacological research. We employed the Semantic Web technologies Resource Description Framework (RDF) and Web Ontology Language (OWL) to generate a Small Molecule Ontology (SMO) that represents concepts and provides unique identifiers for biologically relevant properties of small molecules and their interactions with biomolecules, such as proteins. We instanced SMO using data from three public data sources, i.e., DrugBank, PubChem and UniProt, and converted to RDF triples. Evaluation of SMO by use of predetermined competency questions implemented as SPARQL queries demonstrated that data from chemical and biomolecular data sources were effectively represented and that useful knowledge can be extracted. These results illustrate the potential of Semantic Web technologies in chemical, biological, and pharmacological research and in drug discovery.

1. INTRODUCTION

Many online data resources have been developed to enable researchers to reuse and share data in the life sciences. Major international repositories, notably the European Bioinformatics Institute and the National Center for Biotechnology Information, provide coordinated access to hundreds of data sources particularly in genetic, genomic, and biomolecular science, each of which typically covers different types of data and/or groups of organisms and emphasizes unique sets of properties and features. Researchers thus typically need to integrate data from diverse sources to address complex biological problems. This can be difficult; different data sources may assign the same name to distinct high-level concepts (e.g., "gene" to genomic regions defined by genetic, sequence, comparative, or functional criteria), and these semantic incompatibilities may be further compounded by incongruous naming conventions, idiosyncratic identifiers, and incompatible data formats that impede integration and create opportunities for the propagation of misinformation.

Semantic Web technologies have been proposed as a solution to data integration problems because they present formally defined semantics, make it possible to track data provenance, and support semantically rich knowledge representations.[1] The World Wide Web Consortium (W3C) recommends a suite of Semantic Web enabling technologies, including Extensible Markup Language (XML), Resource Description Framework (RDF) and RDF Schema (RDFS),[2] and the Web Ontology Language (OWL),[3,4] and proposes RDF as the standard model for data interchange on the Web. RDF is modeled as sets of statements, typically referred to as triples, consisting of a subject, a predicate, and an object, where the statement's subject is related to its object through the relationship defined in the predicate. When uniform resource identifiers (URIs) are used as the components of a triple such RDF statements may be used to specify and link related information across the (semantic) Web.

Recently many research groups have endeavored to integrate data effectively from multiple resources in specific domains such as pharmacogenomics,[5] biomedicine,[6] and neuroscience[7] and to develop collaboration frameworks,[8] using Semantic Web technologies. The Bio2RDF project[9] is a more general data integration system to generate RDF triples from publicly available bioinformatics databases and to link those triples through the Bio2RDF URIs.

Low molecular weight chemical compounds of synthetic or natural origin ("small molecules") present many important problems in their own right and interact with biomolecules in multiple ways, including as substrates, allosteric effectors, cofactors, and other ligands. Integrating and sharing comprehensive data describing their structures, measured or computed physical and biological properties, and interactions is likely to yield a powerful impetus to fundamental and applied research, not least in biological chemistry, pharmacology, and drug discovery.[10] Until recently, few small molecule data sources have been available freely online. Some open, online relational databases are now available,[11] for example the Chemical Entities of Biological Interest (ChEBI) database,[12] ChemBank,[13] and PubChem; but as

* Corresponding author. Telephone: +61-7-3346-2616. Fax: +61-7-3346-2101. E-mail: m.ragan@uq.edu.au.

[†] Institute for Molecular Bioscience, The University of Queensland.

[‡] ARC Centre of Excellence in Bioinformatics, The University of Queensland.

[§] Queensland Facility for Advanced Bioinformatics.

10.1021/ci900461j
Published on Web 04/21/2010

earlier with their biomolecular counterparts, integrating these into broader frameworks is impeded both by lack of a semantic framework and by technical issues of knowledge and data representation.[14]

Knowledge representation for small molecule attributes can build on existing naming conventions, standards, controlled vocabularies, ontologies, and conventions for machine-readable chemical data formats, many of them complementary and some in wide use. International Union of Pure and Applied Chemistry (IUPAC; http://www.iupac.org/) names contain structural information and are unique chemical identifers. IUPAC names are to be preferred to common or generic names, which can be imprecise and may not be unique. Medical Subject Headings (MeSH; http://www.nlm.nih.gov/mesh/) and KEGG BRITE (http://www.genome.jp/kegg/brite.html) are controlled vocabularies formed as hierarchical lists to address problems of homographs and synonyms in specific domains and also to cover chemical concepts (chemicals and drugs, and compounds and reactions, respectively). Relevant ontologies are also available, including the Chemical Ontology,[15] ChEBI Ontology,[16] and Functional Group Ontology[17] written in the Open Biomedical Ontologies (OBO) format. While these ontologies capture and represent much valuable and detailed knowledge about chemical space, they do not yet bridge the chemical/biological divide, and important information about biological targets and pathways, along with associated biomolecular domain knowledge, is missing from their respective knowledge models.

A variety of machine-readable formats for chemical data exist and may be used to facilitate adaptation of such data to the Semantic Web. The International Chemical Identifier (InChI), Simplified Molecular Input Line Entry Specification (SMILES), Molfile, Structure-Data Format (SDF), Chemical Mark-up Language (CML), and Protein Data Bank (PDB) formats may all be used to identify or represent information about chemical entities. For example, the CombeChem project, which has been developed to provide a rich set of annotations and flexibility in the sharing and storage of chemical information using Semantic Web technology, has adopted the InChI string as the shared unique identifier of chemical entities.[18,19]

Here we follow an established approach to develop a new ontology that represents key concepts and biologically relevant attributes of small molecule entities. We use RDF, RDFS, XML, and OWL to implement a Small Molecule Ontology (SMO) that represents information about small molecules and their biological targets in a way that is flexible and extensible, yet standard, precise, and formal. We have used OWL Description Logics (OWL-DL)[20,21] as the ontology implementation language to make use of its richly expressive language for describing data and to render these data compatible with emerging standards of the Semantic Web. We present an RDF model of small molecule data, integrating naming conventions and physical attributes of the small molecules themselves as well as information about biomolecules, such as proteins, with which they interact. We create instance data for our ontology by integrating data from public chemical, small molecule, drug activity, and protein resources. We then use the Simple Protocol And RDF Query Language (SPARQL),[22] an RDF query language for extracting information from RDF graphs, to write queries that retrieve novel, useful information from the RDF graph of the instanced data.

By analogy with the application of ontologies in the life sciences and other fields of endeavor, SMO has much potential to facilitate the exploration of small molecule space and to support the discovery of drug candidates and other useful small molecule entities, by providing Semantic Web-assisted integration of biochemical (and chemical) databases with the purpose of supporting the extraction of relationships and new information, including via the application of machine inference. Ontologies, however, are not static entities but instead evolve within the corresponding communities; we present SMO in this spirit.

2. METHODS

2.1. Ontology Development Methodology. A number of ontology development methodologies have been reported in the literature.[23−26] Broadly these approaches can be classified as top-down, middle-out, or bottom-up methods and result in different levels of resolution for the concepts identified through these approaches.[27] We have adopted a middle-out approach in this work to balance granularity of concepts with size and computability.

To develop SMO we followed a methodology based on that of Gruninger and Fox[23,28,29] that includes progressive steps of requirements specification, knowledge acquisition, and implementation and testing and evaluation of the ontology. The strong point of this methodology is to provide a high degree of formality, as it transforms informal competency questions into a computable model expressed in logic. The methodology also includes logical and practical evaluation based on these competency questions. Methodologies for ontology development have been reviewed in detail elsewhere.[30,31]

2.2. Requirements Specification and Knowledge Acquisition. Technical requirements and high-level domain space for the ontology were initially specified. The ontology should: (i) use Semantic Web technologies and public domain data; (ii) support queries and inference; (iii) unify knowledge about small molecules with knowledge about their biological targets; and (iv) where possible, reuse existing repositories of biological and chemical knowledge.

To specify the conceptual coverage required for the ontology, we identified a set of competency questions of different degrees of complexity that represent the kinds of domain space questions that SMO should cover. These questions are then decomposed to identify key concepts and relationships required for the ontology (Table 1). Resources in the knowledge domain are then reviewed to determine the availability of data required to instantiate those concepts. Thus the motivation and the requirements for use of the ontology guide the ontology design and knowledge acquisition process; key concepts and relationships identified from the defined competency questions specify the core classes and the properties that must be included in the ontology in order for it to express adequately the types of constructs indicated by the competency questions. The organization of these concepts and their relationships determine how expressive our data model must be to represent the required data in RDF. Competency questions were then reserved for use

Table 1. Competency Questions and the Corresponding Class Expression

	Competency questions	Concepts	Relationships
Basic	Find **structural** and **identification information** for a **small molecule** of interest	Structural information; Identification information; Small molecule	Small molecule has structure; Small molecule has identification information
	Find all **physical properties** for a **small molecule** of interest	Physical property; Small molecule	Small molecule has physical properties
	Find all **small molecules** which **target** a **protein** of interest	Protein; Target; Small molecule	Small molecule *targets* protein
Complex	Find the names and **subcellular locations** of **proteins** that are **targeted** by a specific **small molecule**	Subcellular location; Protein; Target; Small molecule	Small molecule targets protein; Protein has identifying information name; Protein localized to subcellular location
	For a given **protein**, find its **subcellular locations** and infer more general location information based on **Gene Ontology**, and find **small molecules** which **target** the **protein**	Protein; Subcellular location; Gene Ontology; Small molecule; Target	Protein localized to subcellular location; Small molecule targets protein
	Find **drug-like small molecules** with Lipinski's Rule of Five	Small molecule; Drug-like small molecule	Small molecule *has attribute* drug-likeness
	Find all **proteins targeted** by a specific **small molecule**, and identify **pathways associated with** those **proteins**	Protein; Target; Small molecule; Pathway; Protein	Protein *part of* pathway
	Find all **small molecules** that **target proteins** located in a specific location according to **Gene Ontology**	Small molecule; Target; Protein; Gene Ontology	Small molecule targets protein; Protein localized to subcellular location
	Find all **proteins** that are **associated with** a specific KEGG **pathway**, and find the **small molecules** that **target** these **proteins**	Protein; Pathway; Small molecule; Target; Protein	Protein part of pathway; Small molecule targets protein

in testing and evaluation of the resulting ontology (see Section 2.4).

We adopt an ontology−integration strategy to maximize the use of existing ontologies relevant to the domain of our SMO, following the approach of Pinto and Martins.[32,33] Briefly, we consider ontology integration from the initial stages of our ontology design process. We first chose candidate ontologies for consideration based on their coverage of concepts we identify as important through the decomposition of competency questions (above). After identifying ontologies with relevant conceptual coverage, we then assess the integration operations required to reuse the ontology, apply those operations, and evaluate the resulting ontology (see Section 2.4).

2.3. Implementation. We used the ontology management application framework Protégé 3.4 (http://protege.stanford.edu/) to support the design of the schema for our small molecule ontology.[34] We developed our ontology in OWL-DL based on RDF triples.

2.3.1. Ontology Reuse through Integration. We make use of existing ontologies where possible to capture concepts and relationships that are required for our ontology. Specifically, we use elements from the BioPAX Level2 (Biological Pathway Exchange, http://biopax.org/) ontology that describe physical entities, such as proteins, small molecules, and pathways, and the gene ontology (GO) to describe the functions and locations of gene products.

The BioPAX ontology is written in the target language for our ontology (OWL-DL). BioPAX not only covers metabolic pathways but also supports molecular interactions and post-translational modifications of proteins. BioPAX has two useful classes: physicalEntity (consisting of subclasses small molecule, RNA, DNA, and protein) and pathway, and we use these terms as they are originally defined. BioPAX also has well-developed structures supporting the capture of provenance, metadata, and cross-references, which we reuse for this purpose in our ontology. Terms reused from BioPAX are identified by the namespace prefix bp. Integration operations for BioPAX include a whole ontology import operation (using OWL import statements), and operations on the constituents of the ontology—namely the definition of sets of object properties to integrate classes from BioPAX with our knowledge model.

GO is the most widely used biological ontology and contains three hierarchies of terms describing the biological processes, molecular functions, and cellular components of gene products. GO is natively structured in the OBO format, so we first convert the ontology to OWL, creating a class hierarchy defined by the is_a relationships of GO. Because we restrict our ontology to OWL-DL (to ensure computability), classes may not themselves be used as instances. Therefore an established integration and instancing approach is applied.[35] In brief, we first create an OWL version of the GO OBO file in which each term is represented as a class, and the parent-child relationships modeled in GO as transitive is_a relations are recorded using the owl:subclass relationship. Next, an instance of each class is created to serve as the value of properties in our ontology. Finally, we import the resulting ontology into the SMO using OWL import statements in our ontology, ensuring DL language compatibility, which would be violated by the use of classes as instances. Here, we present the integration of the smaller of the three GO hierarchies, the cellular component (CC) ontology. As with BioPAX, we create a set of custom properties to integrate the concepts present in the GO with concepts in our knowledge model.

2.3.2. Creation of New Classes and Relationships. The majority of nonchemical classes identified through decom-

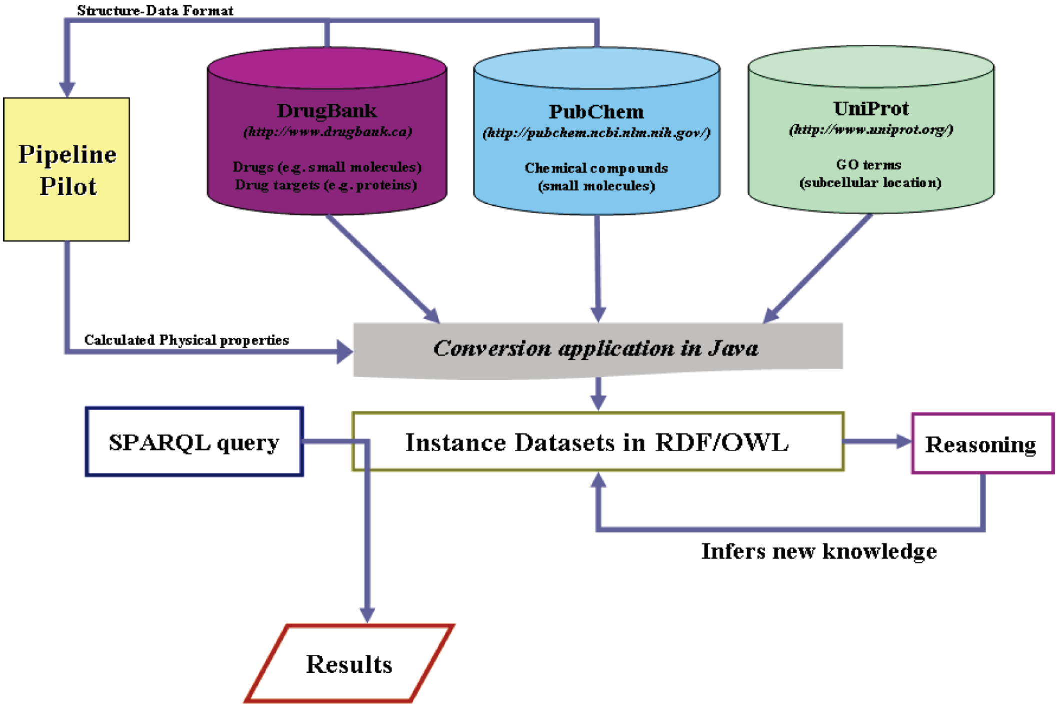
Structure-Data Format
Pipeline Pilot
DrugBank
(http://www.drugbank.ca)
Drugs (e.g. small molecules)
Drug targets (e.g. proteins)
PubChem
(http://pubchem.ncbi.nlm.nih.gov/)
Chemical compounds
(small molecules)
UniProt
(http://www.uniprot.org/)
GO terms
(subcellular location)
Calculated Physical properties
Conversion application in Java
SPARQL query
Instance Datasets in RDF/OWL
Reasoning
Infers new knowledge
Results

Figure 1. Workflow for the creation of instance data for SMO.

position of competency questions are covered by the BioPAX ontology; however, BioPAX lacks detailed reference to the chemical properties of small molecules and fails to give complete conceptual coverage over our application domain. Therefore, we create a blank node _Chem_PhysicalProperty to collect the values of a set of data-type properties that organize information about the physical properties of small molecules. Attributes, such as molecular weight, the number and types of atoms, and relative solubility, potentially informative on solubility and permeability of small molecules, are stored as the values of these data-type properties, and accessed through the blank node.

2.3.3. Population of the Ontology with Instances. To populate our SMO with instances, we first identified publicly available chemical, small molecule, protein, and pathway databases that contain data corresponding to the classes of things present in our ontology (Figure 1). We check that available public data adequately cover the conceptual space of our ontology and identify any missing data. For example, to allow inference of drug-like small molecules among our instance data by use of rule sets for drug-likeness, we need to add physical attributes as properties of small molecules. However, DrugBank and PubChem do not explicitly capture all of the required physical attributes. We, therefore, built a workflow in Pipeline Pilot (http://accelrys.com/products/scitegic/) to calculate missing physical properties of small molecules from the SDF files available for small molecules from DrugBank and PubChem. We use Pipeline Pilot components (SD Reader for input data, PilotScript for calculating physical properties, and Excel Writer for output) to calculate the number of total atoms, heavy atoms, rings, rotatable bonds, hydrogen-bond donors and acceptors, net charge, polar surface area, molar refractivity, and log *P* (octanol−water partition coefficient). Where physical properties were calculated using Pipeline Pilot, the provenance of these data is recorded by setting the value of the *hasDataSource* object property to “PipelinePilot”.

To create instance data from other resources or from privately held data repositories, the Java conversion layer would require customization to deal with any alternative data formats. SDF files may be required for the calculation of physical properties.

DrugBank is a database containing drug data, such as small molecule, pharmacological, and pharmaceutical entities, and drug targets, such as proteins and pathways. It currently contains 4765 nonredundant drug entries which have been approved in North America, Europe, and Asia, and 3037 drug targets.[36,37] We obtained small molecule data for property descriptions from the flat-file text of DrugBank, while other chemical compounds lacking target annotation (i.e., small molecules not known to be drugs) were collected from PubChem. We generate an instance data set of 1000 small molecules for the evaluation of our ontology and demonstrate its data integration capacity by selecting 500 small molecules, each from the DrugBank and PubChem data sets, and by integrating data on protein targets from UniProtKB/Swiss-Prot according to the accession numbers retrieved from DrugBank.

We programmed a conversion application in Java to write the data downloaded from these heterogeneous source databases into RDF-XML instances of our ontology (Figure 1).

2.4. Evaluation. We evaluate our ontology to ensure that it makes technically correct use of Semantic Web technologies, is consistent, satisfies our initial requirements, and supports our competency questions. Here we used the Pellet (http://clarkparsia.com/pellet) command-line interface, an open-source OWL reasoner at JAVA API level, to check for syntactic inconsistencies in our ontology. An ontology is consistent if it is not possible to get contradictory results, given validly defined input.[38] The second evaluation strategy makes use of predefined competency questions.[23] We implemented a set of competency questions (Table 1) as SPARQL queries. SPARQL is a query language for extract-

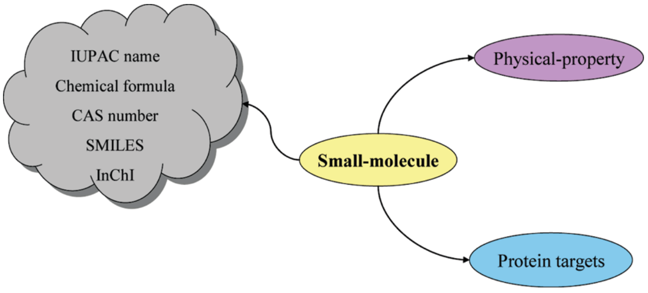

Figure 2. High-level organization of SMO.

ing information from RDF graphs, so a query statement in SPARQL likewise consists of a triple (concept−relation−concept). To write and execute these queries, we used the SPARQL query interface in Protégé 3.4. The Pellet reasoner,[39] combined with Jena (http://jena.sourceforge.net/), is also able to execute SPARQL queries.

We use the Cytoscape plugin RDFScape (http://www.bioinformatics.org/rdfscape/wiki/) to visualize the results of SPARQL queries across our demonstration data set. This plugin, combined with Cytoscape, offers a flexible way to query, visualize, and reason ontological knowledge on ontologies represented OWL or RDF within Cytoscape (http://www.cytoscape.org/). Through the use of queries, a subset or all of the results can be browsed as a graphical network into Cytoscape. An interactive browsing system of RDFScape allows the users to choose one of the menus by right selecting a node as object or subject and then to extend the addition of the relative information to the network.[40]

3. RESULTS

We intend SMO to be a repository of relevant concepts for both small molecule data and target interaction data, usefully describing the chemical and biological attributes of small molecules, supporting integrated discovery and reasoning across small molecules and their biological interaction partners, and facilitating the discovery of drug candidates and other useful small molecules.

3.1. Competency Questions. Competency questions were defined as part of the initial requirements analysis for the SMO. Competency questions (Table 1) capture the context in which we envisage the ontology to be used and serve to identify the conceptual domain that the ontology must cover. Specifically, we seek to provide accurate and flexible retrieval of the identifying information and the physical properties of small molecules as well as their biological targets: proteins and pathways. General concepts were refined by posing specific questions which we have subsequently identified as either basic questions (those that focus on the retrieval of information and the metadata from instances of our ontology) or complex questions (those that rely on the representation of relationships between concepts or that require inference based on the logical consequences of the knowledge model represented in our ontology). Competency questions developed to guide the development of the ontology are then later used in the evaluation of the ontology (see Section 3.4).

3.2. Small Molecule Ontology. Figure 2 shows high-level concepts for describing attributes of small molecules in our SMO in which the key concept *small molecule* includes three specified descriptions: naming and structural data as data types, physical attributes, and protein targets as object types.

The ontology is organized as classes based on concepts identified from the questions listed in Table 1. Many classes available in BioPAX ontology and GO are used to describe related concepts, as we mentioned in Section 2.3.1.

We formally depict classes and relationships between classes or classes and properties in SMO as a hierarchical model using a visualization for RDF and OWL, IsaViz (http://www.w3.org/2001/11/IsaViz/) (Figure 3). High-level concepts used to describe small molecules are shown in Figure 2 and expanded in Figure 3, which illustrates classes, properties (defined as object- or data-type) as well as the specified domain and range of each property.

Some object properties are used to describe the more specific relations: *targets* used to identify small molecules and proteins that the small molecules bind to, *localized* used to identify the proteins and their cellular location information, which is annotated by GO terms, and *part_of* defining pathways in which the proteins are involved.

For rich representation of chemical attributes, we designed the class _Chem_PhysicalProperty as a blank node class to include physical properties for small molecule entities. Attributes such as molecular weight, number and types of atoms, and relative solubility are potentially informative on solubility and permeability of small molecules and as such are used to assess the drug-likeness of small molecules.

In order to describe the new classes we included above, we needed to add new properties such as hasDataSouce, hasPhysicalProperty, and hasTargets for small molecule entities and hasCellularLocation and hasPathway for proteins. The object property hasPhysicalProperty, for example, has bp:smallMolecule as its domain and smo:_Chem_PhysicalProperty as its range, so it defines the relationship between these two classes. Also we created hasDataSource to represent the origin resources of instance data for small molecules and proteins.

3.3. Instance Data in SMO. We constructed a demonstration data set by converting data from DrugBank and PubChem (see Section 2.3). These instances contain 1000 randomly selected small molecules with physical and structural information and, where available, associated biological targets. We included GO annotation specifically to allow us to represent the subcellular locations of small molecule targets as well as pathway references, such as resources and access information.

As a result, we totally converted almost 30 000 RDF triples for 1000 small molecule instances from two databases as shown in Table 2. From Drugbank, 18 745 triples were generated including not only physical and structural attributes but also information of relevant targets with their GO annotations and involved pathway information. PubChem produced 11 000 triples and 22 triples per small molecule with only structural and physical properties, and the total average number of triples for a small molecule is 30.

3.4. Evaluation. We implemented competency questions in the form of SPARQL queries, as semantically correct queries facilitate the evaluation of ontologies. We applied all initial competency questions (Table 1) as queries and reviewed these results to ensure that the SMO returns accurate and meaningful results to these queries (see Supporting Information, File 1). We also tested the ability of the ontology to support inference of new knowledge. For

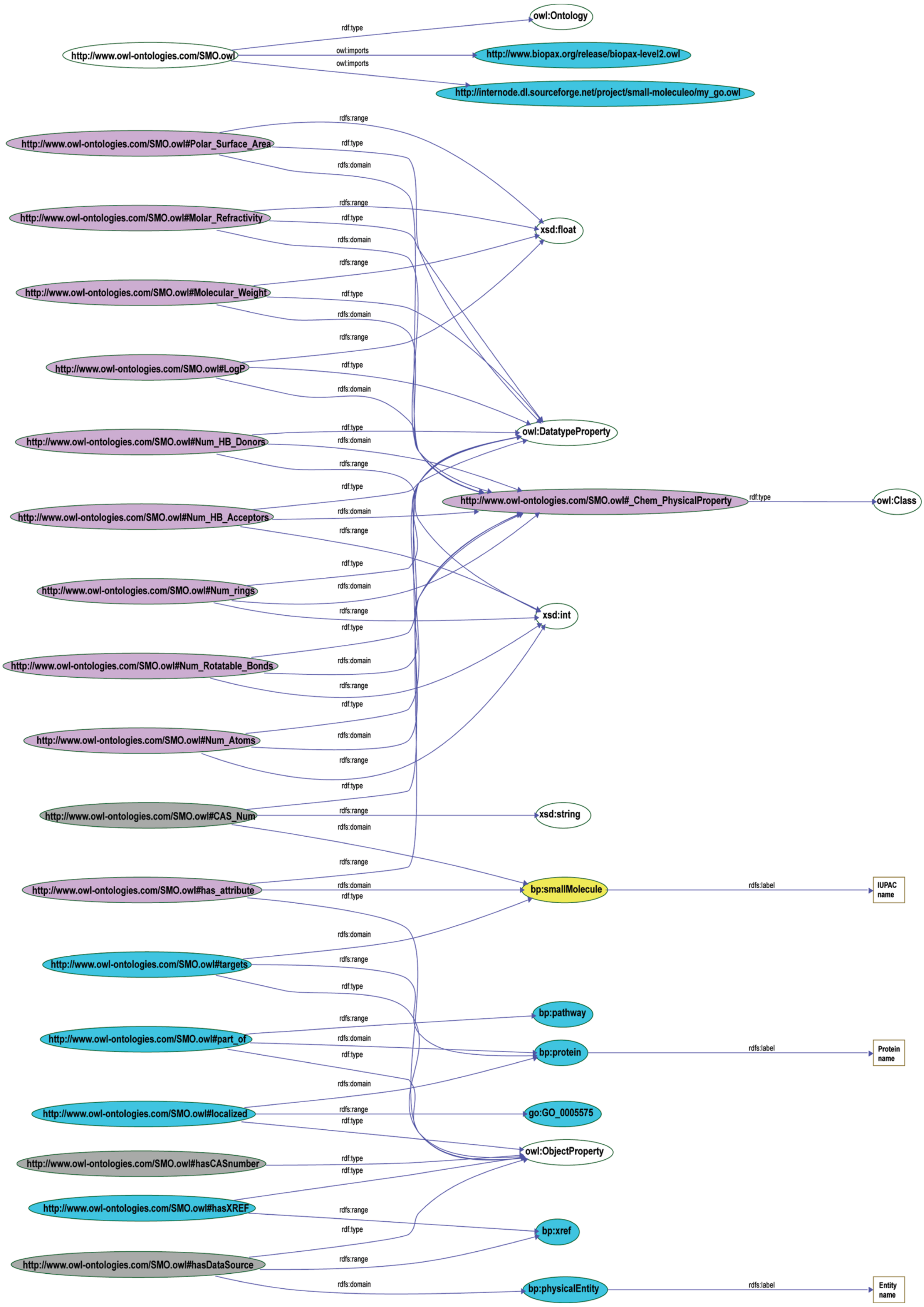
http://www.owl-ontologies.com/SMO.owl
rdf:type
owl:Ontology
owl:imports
http://www.biopax.org/release/biopax-level2.owl
owl:imports
http://internode.dl.sourceforge.net/project/small-moleculeo/my_go.owl
http://www.owl-ontologies.com/SMO.owl#Polar_Surface_Area
rdfs:range
rdf:type
rdfs:domain
http://www.owl-ontologies.com/SMO.owl#Molar_Refractivity
rdfs:range
rdf:type
rdfs:domain
xsd:float
http://www.owl-ontologies.com/SMO.owl#Molecular_Weight
rdfs:range
rdf:type
rdfs:domain
http://www.owl-ontologies.com/SMO.owl#LogP
rdfs:range
rdf:type
rdfs:domain
owl:DatatypeProperty
http://www.owl-ontologies.com/SMO.owl#Num_HB_Donors
rdf:type
rdfs:domain
rdfs:range
http://www.owl-ontologies.com/SMO.owl#_Chem_PhysicalProperty
rdf:type
owl:Class
http://www.owl-ontologies.com/SMO.owl#Num_HB_Acceptors
rdf:type
rdfs:domain
rdfs:range
http://www.owl-ontologies.com/SMO.owl#Num_rings
rdf:type
rdfs:domain
rdfs:range
xsd:int
http://www.owl-ontologies.com/SMO.owl#Num_Rotatable_Bonds
rdf:type
rdfs:domain
rdfs:range
http://www.owl-ontologies.com/SMO.owl#Num_Atoms
rdf:type
rdfs:domain
rdfs:range
http://www.owl-ontologies.com/SMO.owl#CAS_Num
rdf:type
rdfs:range
rdfs:domain
xsd:string
http://www.owl-ontologies.com/SMO.owl#has_attribute
rdfs:range
rdfs:domain
rdf:type
bp:smallMolecule
rdfs:label
IUPAC name
http://www.owl-ontologies.com/SMO.owl#targets
rdfs:domain
rdfs:range
rdf:type
bp:pathway
http://www.owl-ontologies.com/SMO.owl#part_of
rdfs:range
rdfs:domain
rdf:type
bp:protein
rdfs:label
Protein name
http://www.owl-ontologies.com/SMO.owl#localized
rdfs:domain
rdfs:range
rdf:type
go:GO_0005575
owl:ObjectProperty
http://www.owl-ontologies.com/SMO.owl#hasCASnumber
rdf:type
rdf:type
http://www.owl-ontologies.com/SMO.owl#hasXREF
rdfs:range
rdf:type
bp:xref
http://www.owl-ontologies.com/SMO.owl#hasDataSource
rdfs:range
rdfs:domain
bp:physicalEntity
rdfs:label
Entity name

Figure 3. Classes hierarchy of SMO in RDF graph model via IsaViz. We match the same colors for the same concept representing classes and properties with Figure 2. For example, purples indicate the physical attributes of small molecules and include the class of _Chem_PhysicalProperty and the associated nine physical attributes as data-type properties. Classes related to target protein are represented as blue, while naming and structural properties are in gray.

Table 2. Summary of Triple Content for Instance Data Created Using DrugBank and PubChem

	number of RDF triples of instance data set	average number of RDF triples per a molecule
500 data set from DrugBank	18 745	38
500 data set from PubChem	11 000	22
total number of instances	29 745	30

example, the ontology supports reasoning over transitive relations in class hierarchies:

?proteins bp:NAME ‘Estrogen receptor’.

?proteins SMO:localized ?subcellularComponent.

?subcellularComponent rdfs:subClassOf ?superClassOf-Component.

This subset of an example SPARQL query from our competency questions demonstrates the inference of new knowledge: subcellular locations in which small molecules target proteins are retrieved by the object-property SMO: localized, and a property of RDF schemas rdfs:subClassOf is used to infer more general information of the associated subcellular locations for protein targets. The fully executed SPARQL query statement is:

SELECT ?smallMolecule ?subcellularLocation ?superClass

WHERE

{?smallMolecule SMO:targets ?proteins.

?proteins bp:NAME ‘Estrogen receptor’.

?proteins *SMO:localized* ?subcellularLocation.

?subcellularLocation rdf:type ?cellularComponent.

?cellularComponent rdfs:subClassOf ?superClass.}

This query was executed in RDFscape, and the results are visualized as a graph in Figure 4. This example demonstrates how a specific small molecule, tamoxifen, targets the protein ESR1_HUMAN. We retrieve specific cellular locations for the protein and infer more general location information based on the GO classification of cellular components. ESR1_HUMAN is localized in the chromatin_remodeling_complex (GO: 0016585), nucleolus (GO:0005730), cytoplasm (GO:0005737), and plasma_membrane (GO:0005886). The corresponding cellular component superclasses include protein complex (GO:0043234), nuclear part (GO:.044428), intracellular nonmembrane-bounded organelle (GO:0043232), intracellular part (GO:0044424), and membrane (GO:0016020).

Further, we demonstrate the retrieval of implied relationships between small molecules and pathways through target-pathway membership (mereological relations).

?smallMolecule SMO:targets ?proteins.

?proteins SMO:part_of ?pathway.

This subset query retrieves the association of small molecules with pathways through their protein targets. A protein is an element of a pathway, thus a small molecule that targets a protein also targets the pathway in which the protein is involved. Below we represent this question implemented in SPARQL query language, and its result is visualized in Figure 5. This example demonstrates how small molecules that target proteins in a specified pathway can be retrieved. For the pathway query “Glycerolipid metabolism”, the small molecules clofibrate, gemfibrozil, fomepizole, and sortinil are retrieved. Figure 5 demonstrates how the intermediate components, that are the proteins LIPL_HUMAN, ADH1G_HUMAN, and ALDR_HUMAN through which

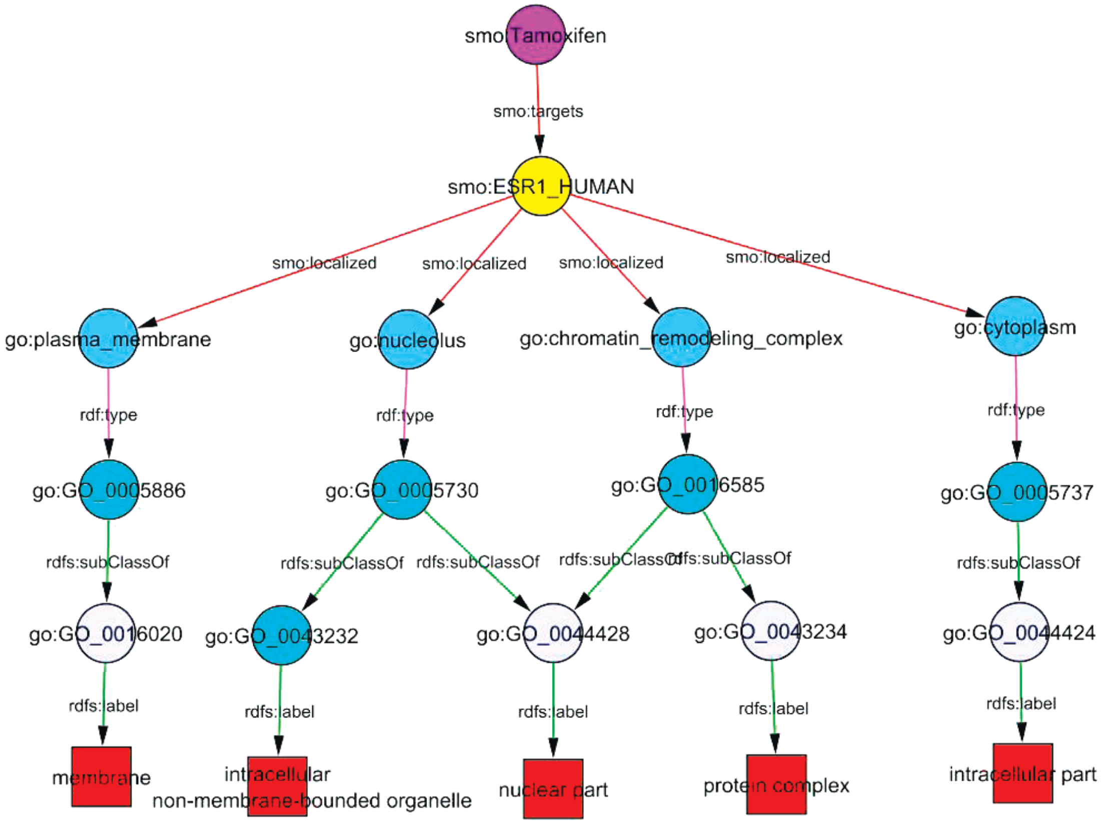

Figure 4. Visualizing the query results in RDFScape: estrogen receptor, the target of small molecule tamoxifen with its annotated subcellular locations and the inference of more-general information via GO; cellular component ‘chromatin remodeling complex (GO:0016585)’ is a subclass of ‘protein complex (GO:0043234)’ and of ‘nuclear part (GO:0044428)’.

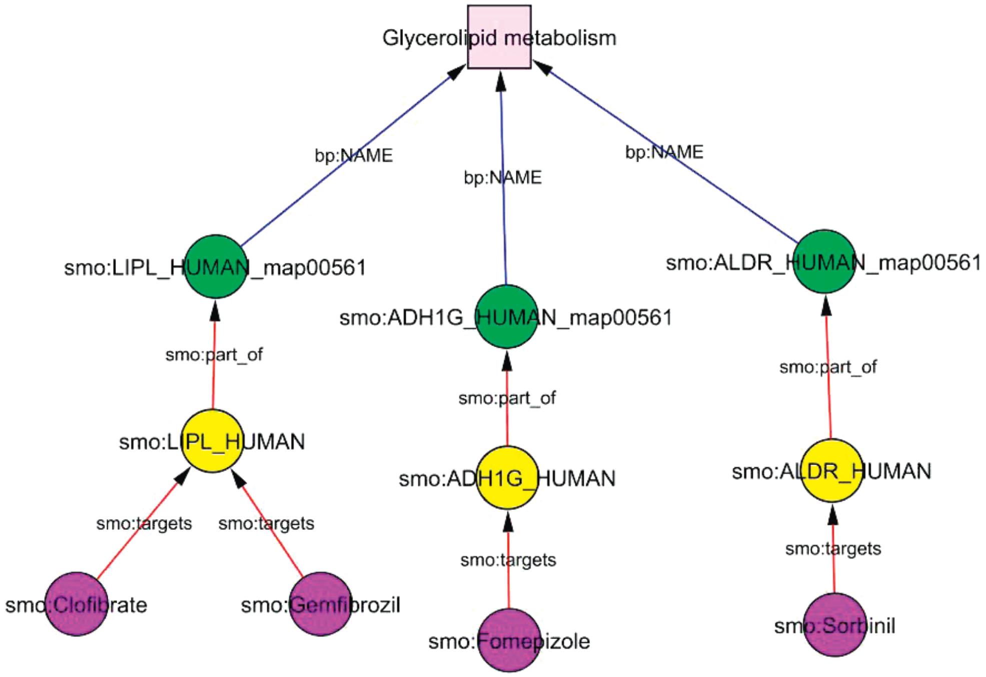

Figure 5. Visualizing the query results in RDFScape: three proteins (e.g., LIPL_HUMAN, ADH1G_HUMAN, and ALDR_HUMAN) in our demonstration data set, which are involved with the 'glycerolipid metabolism' pathway, and small molecules that target these proteins.

these small molecules target the pathway, are also retrieved through exploitation of mereological relationships in the ontology.

```
SELECT ?smallMolecule ?proteins ?pathwayName
WHERE
{?smallMolecule SMO:targets ?proteins.
?proteins SMO:part_of ?pathway.
?pathway bp:NAME ?pathwayName.
FILTER regex (?pathway_name, "Glycerolipid")}
```

3.5. Use and Availability. Our SMO Ontology, in RDF/XML format, is available at http://bioinformatics.org.au/SMO/SMO.owl. The instance data in SMO can be downloaded from http://bioinformatics.org.au/SMO/SMO_1000_instances.owl. The ontology and example data sets are freely available for academic use. The Pipeline Pilot workflow used for calculating physical properties is available at http://bioinformatics.org.au/SMO.

4. DISCUSSION

The development of our SMO builds a knowledge bridge between chemical and pharmacological resources, constructing a unification of knowledge representations of small molecules (or potential drug candidates), proteins as drug targets, and pathways. Unlike existing ontologies of chemical entities, such as Chemical Ontology (CO) or ChEBI ontology, our SMO enables data integration of not only chemical resources but also biological information from heterogeneous sources through Semantic Web languages. We reuse valuable pre-existing knowledge models such as gene ontology (GO), exploiting the detailed class hierarchies of this important biological ontology. We currently use cellular component classifications to demonstrate GO integration, however, the other hierarchies (biological process and molecular function) can be integrated in the same way, thus enabling annotation from these namespaces to be integrated into our knowledge model. Hence, the large knowledge base integrating relevant biological, chemical, and pharmacological information defined by our SMO promises knowledge discovery through the manipulation of and reasoning about high-level concepts and through the access to their detailed instances.

Our work extends the application of Semantic Web technologies in biomolecular science[41−43] and supports the view[44] that these approaches and technologies offer considerable potential in chemical and pharmacological research and drug discovery. We demonstrate that XML, RDF, and OWL-DL can be successfully applied to represent a comprehensive range of concepts pertaining to small molecules, a rich set of their chemical and biological properties, and to biologically relevant interactions between small molecules and biomolecules (e.g., proteins), cellular pathways, and networks. Using a well-described methodology, we generate a novel SMO and evaluate it by use of competency questions. Our evaluation confirms that data from chemical and biomolecular data sources have been effectively represented as RDF triple sets and that useful knowledge can be extracted via SPARQL queries. Similar results were obtained using SWRL rules[45] (results not shown).

The Semantic Web languages RDF/S and OWL offer the potential to extend the representation of evolving sets of data. Using these technologies, it is possible to reuse existing ontologies across different domains, reducing the overheads involved in knowledge acquisition. The RDF/S OWL data model enables inference: as data sets are modeled as a set of relationships between resources, new relationships that may not have been explicitly defined but are the logical implication of the ontology may be inferred to generate new knowledge or insights.

Other benefits previously identified for the use of Semantic Web technologies[19,46] were apparent in this work. The RDF-based data structure is more flexible than a relational data model; data can be stored in RDF format with minimal attention to their specific attributes because attributes are contained in RDF itself. Hence, triples can be stored, additional triples added or removed, and the triple store queried without having to organize tables or to develop a schema and keep it current. The length of InChI strings did not pose a problem under RDF. For example, the InChI string for the small molecule lepirudin is 3337 characters in length and was readily stored as an RDF triple without any modification or compression. RDF is compatible with names that contain nonstandard characters, e.g., in terms from French or German. This flexibility is important for integrating data as diverse and dynamic as those encountered in modern biological and chemical research.

Expansion of SMO into these areas will necessitate the inclusion of additional data sources, e.g., IntAct[47] for interactions between small molecules and proteins and KEGG (LIGAND, DRUG, and COMPOUND). It would likewise be of value to extend the range of small molecules in our instance data by sourcing additional public data sets.

Ideally, ontologies could support chemical and pharmacological research and drug discovery by including concepts of three-dimensional structure as well, e.g., related to structural similarity among small molecules and/or among their biomolecular targets, physical interactions between small molecules and proteins. Ontologies, while good for representing existing data, do not natively support dynamic processing or calculation of properties or relationships. It remains to be seen whether adequate structural detail can be embedded in an ontological representation to retrieve such interactions by queries or by the application of rules such as SWRL, or alternatively whether structural similarities and interactions will need to be precomputed as a data source.

Here we have described a SMO using Semantic Web technologies. SMO provides a semantically rich representation of concepts and unique identifiers for biologically relevant properties of small molecules and their interactions with biomolecules. Our results illustrate that small molecule data and interactions between small molecules and biomolecules, such as proteins, can be effectively represented using Semantic Web technologies, integrated with relevant data resources and used to discover and infer new, useful knowledge relevant to chemical, biochemical, and pharmacological research and to drug discovery.

ACKNOWLEDGMENT

This work was supported by Australian Research Council grant CE0348221 to M.A.R.

Supporting Information Available: The competency questions shown in Table 1 are implemented as SPARQL queries, and the results of these queries are presented as tables following each query. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Baker, C. J. O.; Cheung, K.-H. *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, 1st ed.; Springer: New York, NY, 2007.

(2) Manola, F.; Miller, E. *RDF Primer*; World Wide Web Consortium: Cambridge, MA; Sophia-Antipolis, France; and Tokyo, Japan; http://www.w3.org/TR/2004/REC-rdf-primer-20040210/. Accessed February 28, 2010.

(3) McGuinness, D. L.; Harmelen, F. V. *OWL Web Ontology Language Overview*; World Wide Web Consortium: Cambridge, MA; Sophia-Antipolis, France; and Tokyo, Japan; http://www.w3.org/TR/2004/REC-owl-features-20040210/. Accessed February 28, 2010.

(4) Smith, M. K.; Welty, C.; McGuinness, D. L. *OWL Web Ontology Language Guide*; World Wide Web Consortium: Cambridge, MA; Sophia-Antipolis, France; and Tokyo, Japan; http://www.w3.org/TR/2004/REC-owl-guide-20040210/. Accessed February 28, 2010.

(5) Dumontier, M.; Villanueva-Rosales, N. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings Bioinf.* **2009**, *10* (2), 153–63.

(6) Chen, H.; Ding, L.; Wu, Z.; Yu, T.; Dhanapalan, L.; Chen, J. Y. Semantic web for integrated network analysis in biomedicine. *Briefings Bioinf.* **2009**, *10* (2), 177–92.

(7) Ruttenberg, A.; Rees, J. A.; Samwald, M.; Marshall, M. S. Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings Bioinf.* **2009**, *10* (2), 193–204.

(8) Das, S.; Girard, L.; Green, T.; Weitzman, L.; Lewis-Bowen, A.; Clark, T. Building biomedical web communities using a semantically aware content management system. *Briefings Bioinf.* **2009**, *10* (2), 129–38.

(9) Belleau, F.; Nolin, M. A.; Tourigny, N.; Rigault, P.; Morissette, J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inf.* **2008**, *41* (5), 706–16.

(10) Butcher, E. C.; Berg, E. L.; Kunkel, E. J. Systems biology in drug discovery. *Nat. Biotechnol.* **2004**, *22* (10), 1253–9.

(11) Fullbeck, M.; Michalsky, E.; Dunkel, M.; Preissner, R. Natural products: sources and databases. *Nat. Prod. Rep.* **2006**, *23* (3), 347–56.

(12) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcantara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2008**, *36* (Database issue), D344–50.

(13) Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N. J.; Schreiber, S. L.; Clemons, P. A. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* **2008**, *36* (Database issue), D351–9.

(14) Murray-Rust, P.; Mitchell, J. B.; Rzepa, H. S. Chemistry in bioinformatics. *BMC Bioinformatics* **2005**, *6*, 141.

(15) Feldman, H. J.; Dumontier, M.; Ling, S.; Haider, N.; Hogue, C. W. CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Lett.* **2005**, *579* (21), 4685–91.

(16) Degtyarenko, K.; Hastings, J.; de Matos, P.; Ennis, M. ChEBI: an open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinformatics* **2009**, 14.9.114.9.20, Chapter 14.

(17) Varadwaj, P. K.; Lahiri, T. FGO: A novel ontology for identification of ligand functional group. *Bioinformation* **2007**, *2* (3), 113–8.

(18) Frey, J.; Roure, D. D.; Taylor, K.; Essex, J.; Mills, H.; Zaluska, E. *CombeChem: A Case Strudy in Provenance and Annotation Using the Semantic Web*; Springer Berlin: Heidelberg, Germany, 2006; Vol. 4145/2006, p 270−277.

(19) Taylor, K.; Gledhill, R.; Essex, J.; Frey, J. Bringing Chemical Data onto the Semantic Web. *J. Chem. Inf. Model.* **2006**, *46*, 939–952.

(20) Bechhofer, S.; van Harmelen, F.; Hendler, J.; Horrocks, I.; McGuinness, D. L.; Petel-Schneider, P. F.; Stein, L. A. *OWL Web Ontology Language Reference*; World Wide Web Consortium: Cambridge, MA; Sophia-Antipolis, France; and Tokyo, Japan; http://www.w3.org/TR/2004/REC-owl-ref-20040210/. Accessed February 28, 2010.

(21) Horrocks, I.; Patel-Schneider, P. F.; Harmelen, F. From SHIQ and RDF to OWL: the making of a Web Ontology Language. *J. Web Semant.* **2003**, *1*, 7–26.

(22) Prud'hommeaux, E.; Seaborne, A. *SPARQL Query Language for RDF*; World Wide Web Consortium: Cambridge, MA; Sophia-Antipolis, France; and Tokyo, Japan, 2006.

(23) Gruninger, M.; Fox, M. S. In *Methodology for the design of ontologies and evaluation of ontologies*, Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95 Montreal, 1995; Montreal, 1995.

(24) Fernandez, M.; Gomez-Perez, A.; Juristo, N. In *METHONTOLOGY: From ontological art towards ontological engineering*, AAAI97Spring Symposium Series, Standford, USA, 1997; Stanford, USA, 1997; pp 33−40.

(25) Uschold, M.; Gruninger, M. Ontologies: Principles, methods and applications. *Knowl. Eng. Rev.* **1996**, *11*, 93–136.

(26) Uschold, M.; King, M. In *Towards a methodology for building ontologies*, International Joint Conference on Artificial Intelligence, 1995.

(27) Yu, A. C. Methods in biomedical ontology. *J. Biomed. Inform.* **2006**, *39* (3), 252–66.

(28) Noy, N. F.; McGuinness, D. L. *Ontology Development 101: A Guide to Creating Your First Ontology*; Stanford Knowledge Systems Laboratory Technical Report KSL-01−05; Stanford University: Palo Alto, CA, 2001, *2001*; pp 1−25.

(29) Gomez-Perez, A.; Fernandez-Lopez, M.; Corcho, O. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*, 2nd ed.; Springer: London, 2004.

(30) Corcho, O.; Fernandez-Lopez, M.; Gomez-Perez, A. Methodologies, tools and languages for building ontologies. Where is their meeting point. *Data Knowl. Eng.* **2003**, *46*, 41–64.

(31) Jones, D. M.; Bench-Capon, T. J. M.; Visser, P. R. S. In *Methodologies for ontology development*; Proceedings ITi and KNOWS Conference,15th IFIP World Computer Congress, 1998; Chapman and Hall: London, England, 1998; pp 62−75.

(32) Pinto, H.; Martins, J. In *Reusing ontologies*; AAAI Press: Menlo Park, CA, 2000; pp 77−84.

(33) Pinto, H.; Martins, J. In *A methodology for ontology integration*, International Conference on Knowledge Capture (K-CAP2001), New York, New York, 2001; ACM Press: New York, 2001; pp 131−138.

(34) Rubin, D. L.; Noy, N. F.; Musen, M. A. Protege: a tool for managing and using terminology in radiology applications. *J. Digit. Imaging* **2007**, *20* (Suppl 1), 34–46.

(35) Davis, M. J.; Newman, A.; Khan, I.; Hunter, J.; Ragan, M. A. In *Integrating hierarchical controlled vocabularies with OWL ontology: A case study from the domain of molecular interactions*, Asia-Pacific Bioinformatics Conference, Kyoto, Japan, 2008Brazma, A., Miyano, S., Akutsu, T., Eds.; World Scientific Publishing Company: Kyoto, Japan, 2008; pp 145−154.

(36) Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34* (Database Issue), 668–672.

(37) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheong, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2007**, *36* (Database issue), D901–D906.

(38) Gomez-Perez, F. Evaluation of Ontologies. *Int. J. Intell. Syst.* **2001**, *16*, 391–409.

(39) Sirin, E.; Parsia, B.; Grau, B. C.; Kalyanpur, A.; Karz, Y. Pellet: A practical OWL-DL reasoner. *J. Web Semant.* **2007**, *5*, 51–53.

(40) Splendiani, A. RDFScape: Semantic Web meets systems biology. *BMC Bioinformatics* **2008**, *9*, S6.

(41) Ciccarese, P.; Wu, E.; Wong, G.; Ocana, M.; Kinoshita, J.; Ruttenberg, A.; Clark, T. The SWAN biomedical discourse ontology. *J. Biomed. Inform.* **2008**, *41* (5), 739–51.

(42) Pasquier, C. Biological data integration using Semantic Web technologies. *Biochimie* **2008**, *90* (4), 584–94.

(43) Schulze-Kremer, S. Ontologies for molecular biology and bioinformatics. *In Silico Biol.* **2002**, *2* (3), 179–93.

(44) Villanueva-Rosales, N.; Dumontier, M. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings Bioinf.* **2008**, *10*, 153–163.

(45) Horrocks, I.; Petel-Schneider, P. F.; Boley, H.; Tabet, S.; Grosof, B.; Dean, M. *SWRL: A Semantic Web Rule Language Combining OWL and RuleML;*, 2004; http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/., W3C Member Submission 21 May.

(46) Cheung, K; Yip, K. Y.; Smith, A.; deKnikker, R.; Masiar, A.; Gerstein, M. YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics* **2005**, *21*, 85–96.

(47) Kerrien, S.; Alam-Faruque, Y.; Aranda, B.; Bancarz, I.; Bridge, A.; Derow, C.; Dimmer, E.; Feuermann, M.; Friedrichsen, A.; Huntley, R.; Kohler, C.; Khadake, J.; Leroy, C.; Liban, A.; Lieftink, C.; Montecchi-Palazzi, L.; Orchard, S.; Risse, J.; Robbe, K.; Roechert, B.; Thorneycroft, D.; Zhang, Y.; Apweiler, R.; Hermjakob, H. IntAct−open source resource for molecular interaction data. *Nucleic Acids Res.* **2007**, *35* (Database issue), D561–5.

CI900461J