Journal of Biomedical Informatics 41 (2008) 752–765

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence

Satya S. Sahoo [a,b], Olivier Bodenreider [b,*], Joni L. Rutter [c], Karen J. Skinner [c], Amit P. Sheth [a]

[a] Kno.e.sis Center, Wright State University, Dayton, OH, USA
[b] LHNCBC, National Library of Medicine, 8600 Rockville Pike, MS 3841, Building 38A, Room B1N28U, Bethesda, MD 20894, USA
[c] DBNBR, National Institute on Drug Abuse, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 6 October 2007
Available online 29 February 2008

Keywords:
Semantic web
Semantic mashup
Nicotine dependence
Information integration
Ontologies

ABSTRACT

Objectives: This paper illustrates how Semantic Web technologies (especially RDF, OWL, and SPARQL) can support information integration and make it easy to create semantic mashups (semantically integrated resources). In the context of understanding the genetic basis of nicotine dependence, we integrate gene and pathway information and show how three complex biological queries can be answered by the integrated knowledge base.

Methods: We use an ontology-driven approach to integrate two gene resources (Entrez Gene and HomoloGene) and three pathway resources (KEGG, Reactome and BioCyc), for five organisms, including humans. We created the Entrez Knowledge Model (EKoM), an information model in OWL for the gene resources, and integrated it with the extant BioPAX ontology designed for pathway resources. The integrated schema is populated with data from the pathway resources, publicly available in BioPAX-compatible format, and gene resources for which a population procedure was created. The SPARQL query language is used to formulate queries over the integrated knowledge base to answer the three biological queries.

Results: Simple SPARQL queries could easily identify hub genes, i.e., those genes whose gene products participate in many pathways or interact with many other gene products. The identification of the genes expressed in the brain turned out to be more difficult, due to the lack of a common identification scheme for proteins.

Conclusion: Semantic Web technologies provide a valid framework for information integration in the life sciences. Ontology-driven integration represents a flexible, sustainable and extensible solution to the integration of large volumes of information. Additional resources, which enable the creation of mappings between information sources, are required to compensate for heterogeneity across namespaces.

Resource pagehttp://knoesis.wright.edu/research/lifesci/integration/structured_data/JBI-2008/

1. Introduction

It is estimated that, worldwide, over one billion people smoke tobacco. The detrimental consequences of smoking on health are well known and include coronary heart disease, lung cancer and chronic obstructive pulmonary disease. The heritability of nicotine dependence has long been established and we know that approximately 40–60% of nicotine dependence is due to genetic contributions, while the remainder is largely environmental [1–3]. In the past few years, genome-wide linkage and association studies have identified several candidate genes (e.g., GABAB2, CHRNA4, DDC, BDNF, and COMT.) [4–6]. Saccone et al. identified and screened 449 human genes putatively involved with nicotine dependence [6]. In addition to identifying the genes, it is important to understand their functions and interactions, including their involvement in biological pathways. For example, from a research management perspective, identification of "hub" genes (i.e., genes involved in multiple pathways) can help identify further research efforts.

Complex biological queries generally require the integration of information from several sources. For example, gene information sources, such as Entrez Gene [7], might need to be integrated with pathway information sources, such as KEGG (Kyoto Encyclopedia for Genes and Genomics) [8]. Moreover, comparing results across model organisms requires homology information (provided for example by HomoloGene [9]). These resources, described in detail later in Section 4.4, are generally cross-referenced, which makes it possible for users to navigate among them in web-based environments. Interlinking is not the same as integration, however; and these resources do not support the automatic and high-throughput information processing required for answering complex queries over large amounts of data from heterogeneous sources. An effec-

* Corresponding author. Fax: +1 301 480 3035.
E-mail address: olivier@nlm.nih.gov (O. Bodenreider).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.02.006

tive integration strategy is also critical to support the e-Science paradigm that is characterized by the large volumes of data generated by industrial-scale *in-silico* processes [10].

The first obstacle to integration is the format used for the representation of these information sources. The resources available from the National Center for Biotechnology Information (NCBI) Entrez system, such as Entrez Gene and HomoloGene, are available in multiple formats, including XML. Although XML standardizes the representation of information from a syntactic perspective, it does not make explicit the relations among the various types of entities in a given resource or across resources. In other words, although the XML file for Entrez Gene is machine-processable, it cannot be integrated easily or automatically with other information sources without human intervention. In contrast, the pathway research community has created a common, formal knowledge model called BioPAX [11] to represent biological pathway data. BioPAX also provides an information model for representing those data with formally defined semantics, which includes explicitly modeling the relationships between different pathway entities.

Recent research in Semantic Web technologies has delivered promising results for information integration across heterogeneous knowledge sources [12–15]. In effect, the Semantic Web provides a robust framework that enables the integration, sharing, and reuse of data from multiple sources. Additionally, the use of a representation based on a formal language allows software applications to reason over information. Commonly used Semantic Web technologies include ontology modeling languages such as Web Ontology Language (OWL) [16], data models such as the Resource Description Framework (RDF) [17], the SPARQL query language [18], and OWL reasoners such as Pellet [19] and Racer [20]. Reasoning tools have been successfully applied over knowledge bases to address biological and health care problems [13,21–25].

The objective of this paper is to illustrate how Semantic Web technologies can support information integration and facilitate the creation of semantically integrated resources, called semantic mashups, for gene and pathway information. We show how complex biological queries can be answered by the mashups. More precisely, in the context of understanding the genetic basis of nicotine dependence, we integrate gene and pathway information in order to answer the following queries: which genes participate in a large number of pathways? Which genes (or gene products) interact with each other? Which genes are expressed in the brain?

The rest of the paper is organized as follows. In Section 2, we justify the use of ontology-based integration and summarize relevant work on the use of Semantic Web technologies in biomedicine. In Section 3, we present the ontological framework we created to support the integration. The information sources integrated are presented in Section 4, along with our integration strategy. Three biological queries and the corresponding answers extracted from our knowledge base are explored in Section 5. In Section 6, we discuss the significance of this study, as well as its limitations. Our conclusions are presented in Section 7. Supplementary materials, including the set of SPARQL queries used in this study and the EKoM schema, are available at http://knoesis.wright.edu/research/lifesci/integration/structured_data/JBI-2008/.

2. Background

In this section, we discuss the rationale for ontology-based integration and summarize relevant work on the use of Semantic Web technologies in biomedicine.

2.1. Ontology-based data integration

The traditional approach to integrating gene and pathway information is to create a relational data model that can be used to integrate and store both kinds of data. As the experience of the biological pathway research community with the BioPAX ontology clearly shows, there are many advantages to using an ontology as a knowledge representation model for integrating data from heterogeneous sources [11,22]. One advantage is that the formal semantics of an ontology enable software applications to interpret ontology instance data consistently and reason over them. For example, the entities '*gene*' and '*molecular function*' are represented in the ontology, where they are linked by the relationship '*has_function*.'[1] At the instance level, a particular gene (e.g., Chrna4 in mouse) has a particular function (e.g., 'nicotinic acetylcholine-activated cation-selective channel activity'). This advantage has been discussed in a wide range of application domains including national security [26], geographical information systems [27] and biomedical informatics [28].

Complex biological queries require precisely this kind of reasoning over a large number of instances. Although scientists can easily interpret the connections among entities, they generally are unable to process large amounts of data consistently. Conversely, computers can identify connections in large graphs, but require that relations be explicitly represented. To identify common pathways among homologous genes from the 449 genes putatively involved with nicotine dependence, for example, two types of (instance-level) information need to be extracted from the relevant knowledge bases and processed: homology information and gene-pathway relations. The corresponding types of entities (here, '*gene*' and '*pathway*') and relations ('*homologous_with*' and '*involved_in*') must be represented in the corresponding information model or ontology.

From a theoretical perspective, without a knowledge model associated with the RDF instance data, the discovery of new knowledge through entailment reasoning will be limited. (Entailment reasoning rests on the notion that if formula A entails formula B, then every interpretation satisfying A also satisfies B). Simple RDF interpretation and entailment ignore the ''meaning of any of the names in the graph,'' as described in RDF semantics [29]. Moreover, the W3C RDF semantics recommendation [29] suggests attaching stronger meaning to URI references to gain maximum value from an RDF graph written ''in a particular vocabulary.'' Treating an ontology as a vocabulary, with clearly defined concepts and relationships that are used to 'type' instance values, enables class membership-based entailment reasoning. Creating an RDF graph by using an ontology as the reference knowledge model leads to a ''stronger notion of interpretation and entailment'' [29].

The BioPAX ontology provides a common information model based on RDF/XML syntax for various pathway information sources, including, KEGG, Reactome, and BioCyc. Conversely, the gene information sources available through NCBI's Entrez system, for example, Entrez Gene and HomoloGene, are by design available only in XML format, and no common information model representing their semantics is provided. We therefore created the Entrez Knowledge Model (EKoM) to represent NCBI gene information in a formal semantic model. We then created schema-level mappings between the BioPAX ontology and EKoM, to integrate the two into a single global schema for representing both gene and pathway data. (The details are presented in Section 3, Ontology creation and schema mapping).

Ontology-based data integration, subscribing to the Local As View (LAV) data integration theory [30], not only uses the formal semantics of the ontology language, but is also a scalable and adaptable integration approach. The LAV approach involves the representation of data from original sources in conformance to a

[1] In this paper, we represent ontology concepts in italics and within single quotes (e.g. '*pathway*').

common model or schema such as an ontology. An important aspect of the LAV approach is that it "favors the extensibility of the system" [30], which is critical here, as other data sources may be added in the future.

Another significant feature of ontology-based data integration approach is the use of inference mechanisms for information gain. An ontology is created using a formal language, the Description Logic-based flavor of OWL (OWL-DL) in the case of BioPAX and EKoM, which allows the definition of inference rules that can be interpreted and processed by reasoning tools. Given two genes that interact with one another, for example, we can define an inference rule to assert a new relationship that exists between their respective proteins products (they either bind together or form components of a larger biological pathway).

2.2. Related work

In previous work, we successfully created an RDF representation of the complete Entrez Gene data set [31] by mapping the XML element tags to named relationships. We used XML Path language [32] with eXtensible Stylesheet Language Transformation (XSLT) [33] approach to make the conversion from the native Entrez Gene XML representation to RDF. Subsequently, we integrated this Entrez Gene RDF data with the publicly available Gene Ontology (GO) RDF dataset. Using a set of rules, we showed how phenotypic and genotypic information can easily be linked using RDF [31]. Specifically, we demonstrated the existence of a link between the disease 'congenital muscular dystrophy' and GO molecular function 'glycosyltransferase.'

There is a growing body of research related to the application of Semantic Web technologies to the life sciences domain [12,14,15,34–36]; this section discusses some of these efforts. Chabalier et al. [37] describes work involving the classification of diseases along physio-pathological classes and the identification of taxonomic relations between diseases using KEGG pathway data and GO annotations. Ruttenberg et al. [13] presents an overview of work by the World Wide Web Consortium (W3C) Health Care and Life Sciences Interest Group (HCLSIG) on the use of Semantic Web technologies toward achieving the vision of "Translational Medicine." Many interesting projects were discussed at the World Wide Web (WWW) conference at Banff, Canada, in 2007, including the use of Semantic Web technologies for mining disease-causing genes through integration of genome–phenome data [38] and a semantic mashup created by the HCLSIG to aid neurosciences researchers [21].

These projects are similar to the work described in this paper in that they highlight the use of Semantic Web technologies such as RDF, OWL and SPARQL to achieve relevant biomedical objectives using data from heterogeneous sources. The work described in this paper is also a natural progression from our previous work [31] and the HCLSIG demonstration [21]. Distinct differences include the use of ontologies as a reference model with associated formal semantics, schema mapping between EKoM and the BioPAX ontology, and rules to reconcile heterogeneous instance bases.

3. Ontological framework

In this section, we give a brief presentation of the BioPAX ontology and discuss the design decisions we made while creating the Entrez Knowledge Model (EKoM). Subsequently, we describe the mapping we created between the EKoM and the BioPAX ontology.

3.1. BioPAX ontology

The BioPAX ontology was created to model biomolecular pathways [39]. There are two BioPAX ontology releases namely level 1, which represents only metabolic pathways, and level 2, which in addition represents molecular interactions, protein post-translational modifications, and the Protein Standards Initiative-Molecular Interactions (PSI-MI, http://www.psidev.info/) [40]. The BioPAX ontology, level 2, used in the work described in this paper, defines pathway data in terms of concepts such as '*interaction,*' '*entity,*' or '*pathway*'(Fig. 1) and relationships between them such as '*pathway_components*' (between '*pathway*' and '*interaction*') and '*participants*' (between '*entity*' and '*interaction*'). The BioPAX ontology (level 2) is modeled using the OWL-DL language with DL expressivity of *ALCHON (D)* and has 40 classes with 33 object and 37 data type properties.

3.2. Entrez knowledge model

Since no formal information model is available for the representation of gene information in the Entrez family of sources (e.g., from Entrez Gene and HomoloGene), we considered the following approaches to creating such a model. We could either create a new ontology to represent gene information from NCBI sources or extend the BioPAX ontology schema to include the concepts and relationships relevant to gene information sources. We chose not to extend the BioPAX ontology, since it was created specifically to model bio-molecular pathways [39]. Our goal in developing the Entrez Knowledge Model (EKoM) is to create a standalone model specific to NCBI gene information sources, and to integrate it with other models such as the BioPAX ontology.

Gene records from Entrez Gene contain information about the gene product(s), the chromosomal location of the genes, the model organisms in which they are found, and the pathways in which these genes are involved. Each record also contains information such as its creation date and current status. The entities represented in EKoM correspond to records in databases (e.g., Entrez Gene records), which we use as a proxy to the corresponding entities in reality (e.g., genes).

EKoM is modeled using the OWL-DL language with *SI* DL expressivity (i.e., it uses concept negation, universal and existential quantification, intersection, and disjunction between concepts as well as inverse role for relations). There currently are 45 classes defined in EKoM, through which we have tried to capture essential

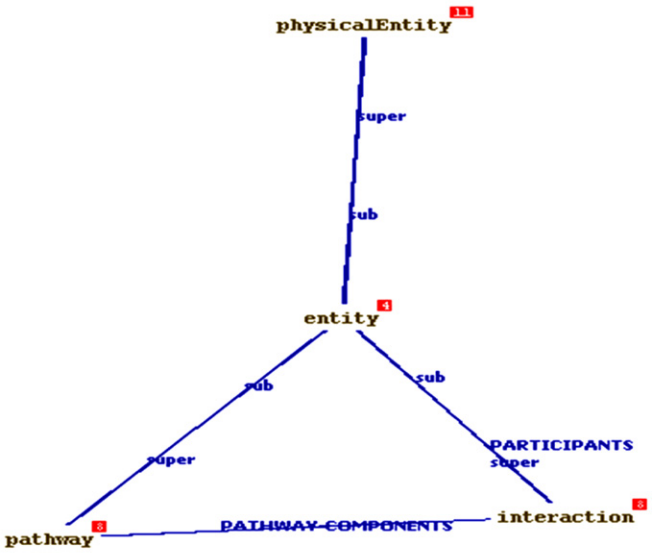

Fig. 1. Top-level BioPAX concepts and relationships (Protégé TGViz plug-in diagram with 1-level fan-out).

data available in Entrez Gene (Fig. 2). There are 11 relationships defined in EKoM as object properties, which link together the classes. These named relationships are either defined in the UMLS Semantic Network [41] or specifically created for EKoM. (The relationships defined in reference ontologies, such as the OBO Relation Ontology [42], were generally too coarse to be useful here.)

Using the example of gene–gene interaction, we describe the modeling approach used for EKoM. Fig. 3 illustrates our approach as applied to a specific gene (CHRNA4). The interaction information from Entrez Gene records includes the "original" gene (CHRNA4) and its gene product (cholinergic receptor, nicotinic, alpha-4 subunit), the "interactant" gene (CHRNB2) and its gene product (neuronal nicotinic acetylcholine receptor beta-2), the textual description of the interaction ("CHRNA4 (alpha-4) interacts with CHRNB2 (beta-2)"), and a reference in the form of a PubMed identifier. We modeled this information using the concepts '*gene*,' '*interaction*,' '*gene_product*,' '*protein_db_identifier*,' and '*reference*.'

3.3. Schema-level integration of EKoM and BioPAX ontology

With the BioPAX ontology, modeling pathway information, and EKoM, modeling gene information in sources from the Entrez system, we have two information models (or schemas) that need to be integrated. We found three potentially similar concepts in EKoM and the BioPAX ontology, namely '*pathway*,' '*protein*,' and '*interaction*.' We chose to reuse the concepts '*pathway*' and '*protein*' in EKoM, as defined in the BioPAX ontology, instead of redefining them in EKoM. In contrast, although the '*interaction*'concept is present in both EKoM and BioPAX, we identified that its meaning was different in the two models. In fact, BioPAX states "*Since ['interaction'] is a highly abstract class in the ontology, instances of the interaction class should never be created. Instead, more specific classes should be used...*" On the other hand, EKoM does not define any subclasses for '*interaction*' and instantiates this class directly. Therefore, the concept '*interaction*' defined in BioPAX ontology was not reused in EKoM.

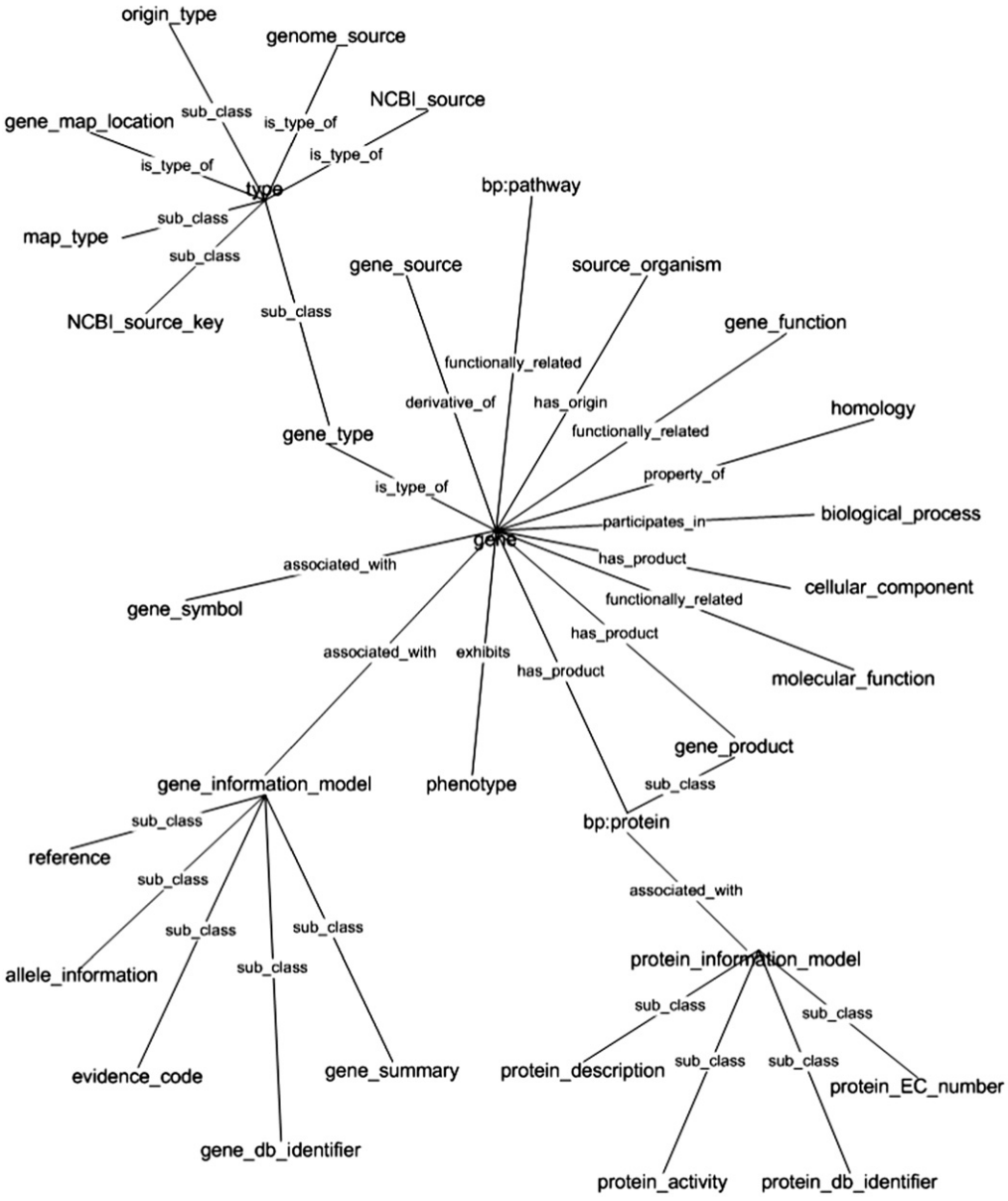

Fig. 2. Top level concepts and relationships of EKoM.

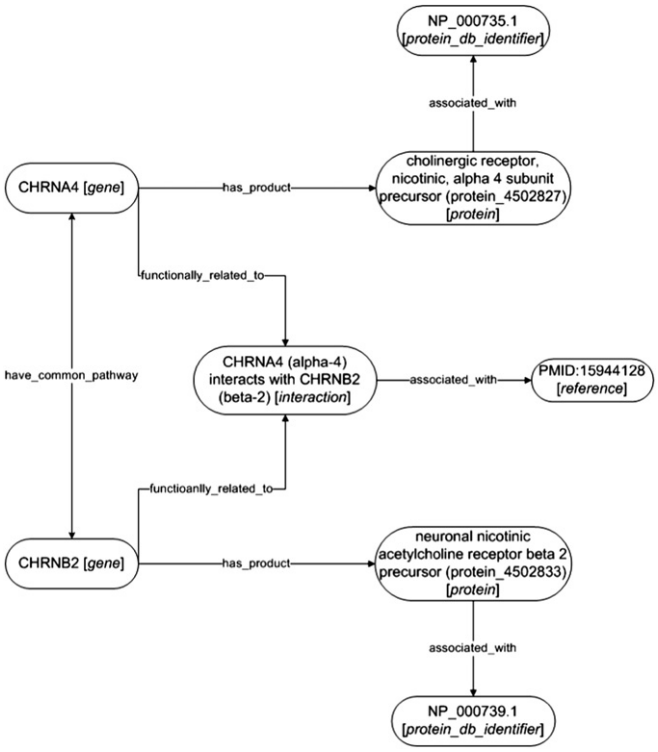

Fig. 3. Interaction between genes modeled in EKoM (using gene 1137 and 1141 as examples).

We integrated the two schemas by importing the BioPAX definition of the concepts in EKoM and created relations between EKoM and BioPAX concepts as appropriate. In practice, a gene-pathway relation is represented as the relation (defined in EKoM) between a gene (EKoM concept) and a pathway (BioPAX concept). For example, the relation *EKoM:gene_6261 → EKoM:functionally_related_to → bp:KEGGpathway_04730* between the gene CHRNB2 (GeneID: 6261) and the pathway Long-term depression (KEGG: 04730). The global schema resulting from the mapping between EKoM and BioPAX provides a formal semantic framework for integrating the data from gene resources and pathway resources at the instance level. Specifically, it is implemented through class membership relations between specific entities (e.g., CHRNB2) in these resources and the corresponding concepts in the information model (e.g. *'gene'*).

3.3.1. Common ontology schema—heterogeneous instance bases

The availability of pathway information from three large sources (KEGG, Reactome and BioCyc) conforming to the BioPAX ontology offers a critical advantage in building a semantic pathway knowledge repository. The assumption, regarding the three instance datasets, is that the common ontology schema will enable their automatic and seamless integration [22]. In practice, however, the BioPAX ontology (used as knowledge model for the three resources) seems to be interpreted slightly differently by each data source provider. In fact, the semantics of these class properties varies across resources, resulting in heterogeneous instance bases despite the common ontology schema. More specifically, the instantiation of BioPAX ontology differs in the following aspects:

- In KEGG, the URI for '*pathway*' instances is based on a unique alpha-numeric identifier. The value of the '*SHORT-NAME*' property is the KEGG identifier of the pathway. The value of the '*NAME*' property is the textual description of the pathway. Both values are typed as a 'XML schema string' (http://www.w3.org/2001/XMLSchema#string).
- In Reactome, the URI for '*pathway*' instances is based on the textual description of the pathway. The values of both '*SHORT-NAME*' and '*NAME*' properties are textual descriptions of the pathway. The pathway identifier is associated with the '*XREF*' property.
- In BioCyc, the URI for '*pathway*' instances is based on the BioCyc identifier. The value of the '*NAME*' property is the textual description of the pathway. No other property from the '*pathway*' concept is used.

As illustrated above, the instantiation of the '*pathway*' concept in the three pathway resources under investigation differs in subtle but significant ways. In practice, two major types of issues are identified. First, '*pathway*' instances cannot be compared on the basis of their URIs. In addition, the semantics of the properties of the '*pathway*' concept in BioPAX (e.g., '*SHORT-NAME*') differs across resources. As a consequence, the instance bases for the three resources are heterogeneous, and '*pathway*' instances cannot be easily compared on the basis of the value of these properties.

These heterogeneous instance bases can be potentially reconciled by using related knowledge such as relationships and values associated with each '*pathway*' instance. The Pathway Knowledge Base (PKB) [43] discusses the integration of the three BioPAX conformant data sources using Jena [44] to create RDF objects to create a unified store. To support querying across the three data sources, PKB preprocesses the data for syntactic reconciliation including uniformly converting all BioPAX level 1 references to level 2 and use of a standardized namespace (http://pkb.stanford.edu). In our work, we focused on semantic reconciliation that is partially based on syntactic reconciliation (discussed in the next section).

3.3.2. Reconciling heterogeneity among instances

The heterogeneity among pathway instances is not only syntactic (e.g., different format for the identifiers), but also semantic. For example, as described in the previous section, the identifier for a pathway instance in Reactome is the textual description of the pathway, whereas in KEGG it is a unique alpha-numeric value. We used additional knowledge associated with a pathway instance to assess whether two instances are semantically identical or not. This additional knowledge comes from named relationships, for example '*bp:SHORT-NAME*' and '*bp:XREF,*' linking the pathway instance to other entities such as database identifiers and textual descriptions.

For example, as illustrated in Fig. 4, although the instances for calcium signaling pathway are distinct in Entrez Gene and KEGG, we observed that they share the same value (hsa04020) for the '*SHORT-NAME*' property. We created a rule to assert the equivalence between the corresponding instances: if two instances from Entrez Gene and KEGG share the same value for the '*SHORT-NAME*' property, then they must be considered as one. Technically, we assert an '*owl:sameAs*' relation between the two instances, so that a reasoner can interpret them as being semantically identical. Similarly, the '*pathway*' instances from Entrez Gene and Reactome that share the same value for the '*XREF*' property values are asserted to be identical.

4. Materials and architecture

The hypothesis underlying this study is that a mashup of gene and pathway resources created with Semantic Web technologies will help answer complex biological queries related to the genetic basis of nicotine dependence. The primary gene resource to be

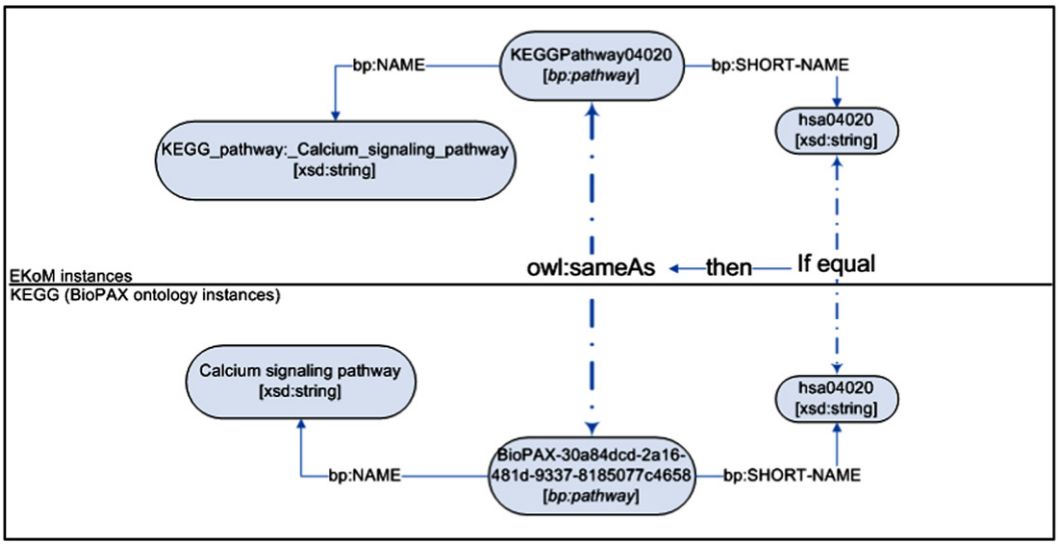

Fig. 4. Reconciling KEGG and EKoM pathway instances through a rule.

integrated is Entrez Gene, while Entrez HomoloGene is used to identify homologous genes in various model organisms. In addition, three pathway information sources—KEGG, Reactome, and BioCyc—are also integrated in the mashup. While the three pathway sources are already available in RDF/XML and conform to the BioPAX ontology schema, the gene resources Entrez Gene and HomoloGene, available in XML, first need to be converted to RDF, conforming to EKoM, the information model we created for these resources. Fig. 5 describes the procedure for creating the knowledge base and the overall architecture of the system.

4.1. Mapping nicotine dependence genes to Entrez Gene

The list of human genes described in [6] (henceforth referred to as the original set of genes) consists primarily of gene names and gene symbols. In addition, chromosomal location is provided for most genes and a short textual description is provided for some genes. Using tools from the eUtils family (EFetch and ESearch) [45], we retrieved Entrez Gene records corresponding to the gene symbols and validated them against 'Gene Name,' 'Organism,' and 'Chromosomal location' fields from the record. For example, the gene symbol SNAP25 maps unambiguously (when restricted to human genes) to the gene identified by GeneID: 6616 in Entrez Gene. The mapping was straightforward for 80% of the genes. In 48 cases, however, multiple records or, more rarely, no records were retrieved from the gene symbol. Ambiguous symbols were disambiguated manually using additional information such as the gene name or chromosomal location. For example, TF mapped to both TF and F3 (for which TF is an alias), and was subsequently disambiguated to TF (GeneID: 7018) using the name "transferrin." When no record was found for the gene symbol, the gene name was used instead of the symbol to query Entrez Gene. For example, the symbol CALCYON did not map to any human genes, but the corresponding name in the original set (D1 Dopamine Receptor-Interacting Protein) mapped to the gene DRD1IP (GeneID: 50632). At the end of this process, a unique Entrez Gene record was found for each of the 449 genes in the list.

4.2. Identifying homologous genes

HomoloGene contains homology data for several completely sequenced eukaryotic organisms [9]. Entrez Gene records contain HomoloGene identifiers that can be used as pointers to homologous genes. In addition to *Homo sapiens* (taxId: 9606), four model organisms were considered in this study, because they exhibit biological processes known to be related to nicotine dependence. The model organisms under investigation are *Mus musculus* (taxId: 10090), *Caenorhabditis elegans* (taxId: 6239), *Danio rerio* (taxId: 7955) and *Drosophila melanogaster* (taxId: 7227). Beginning with a record in Entrez Gene (e.g., ALDH2, GeneID: 217), a link is found to record 55480 in HomoloGene, from which the Entrez Gene IDs for homologous genes can be extracted. For example, this HomoloGene record identifies aldh2a (GeneID: 393462) in zebrafish. Here again, the process of identifying homologous genes from HomoloGene is completely automated through the use of EFetch. A total of 1,401 gene records were extracted from Entrez Gene. In addition to the 449 gene records for *H. sapiens*, we retrieved the records for homologous genes in the following model organisms: *M. musculus* (381), *C. elegans* (99), *D. rerio* (364) and *D. melanogaster* (108).

4.3. Acquiring gene information

Resources from the Entrez family, including Entrez Gene and HomoloGene, are made available by NCBI in XML. However, we use RDF/XML for the representation of our integrated gene-pathway resource. Therefore, we need to convert XML records from Entrez Gene and HomoloGene to RDF. Moreover, in order to be able to reason over the RDF store, we require the RDF data to conform to the Entrez Knowledge Model (EKoM) we created for this purpose. In other words, entities from the RDF store become instances of the classes and relationships in EKoM schema.

In previous work [31,46], we developed a method for converting Entrez Gene records from XML to RDF, based on XPath and XLST stylesheet transformation. The mapping (created manually) between element tags in XML and properties in RDF is recorded as a set of transformation rules in the stylesheet. This process transforms a relation, implicitly represented in Entrez Gene (e.g., between a given gene and its gene product), into a RDF triple in which this relation is made explicit. Our current works expands this procedure by creating class-membership relations between the instances in RDF and classes from EKoM. For example, the relation '*has_product*' between the gene CHRNA4 (GeneID: 1137) (with XML element tag '<Gene-track_geneid>') and the protein *cholinergic receptor, nicotinic, alpha 4 subunit precursor* (GI: 4502827) (with XML element tag '<Prot-ref_name_E>') is made explicit and transformed into a RDF triple *GeneID: 1137 → has_product → GI:*

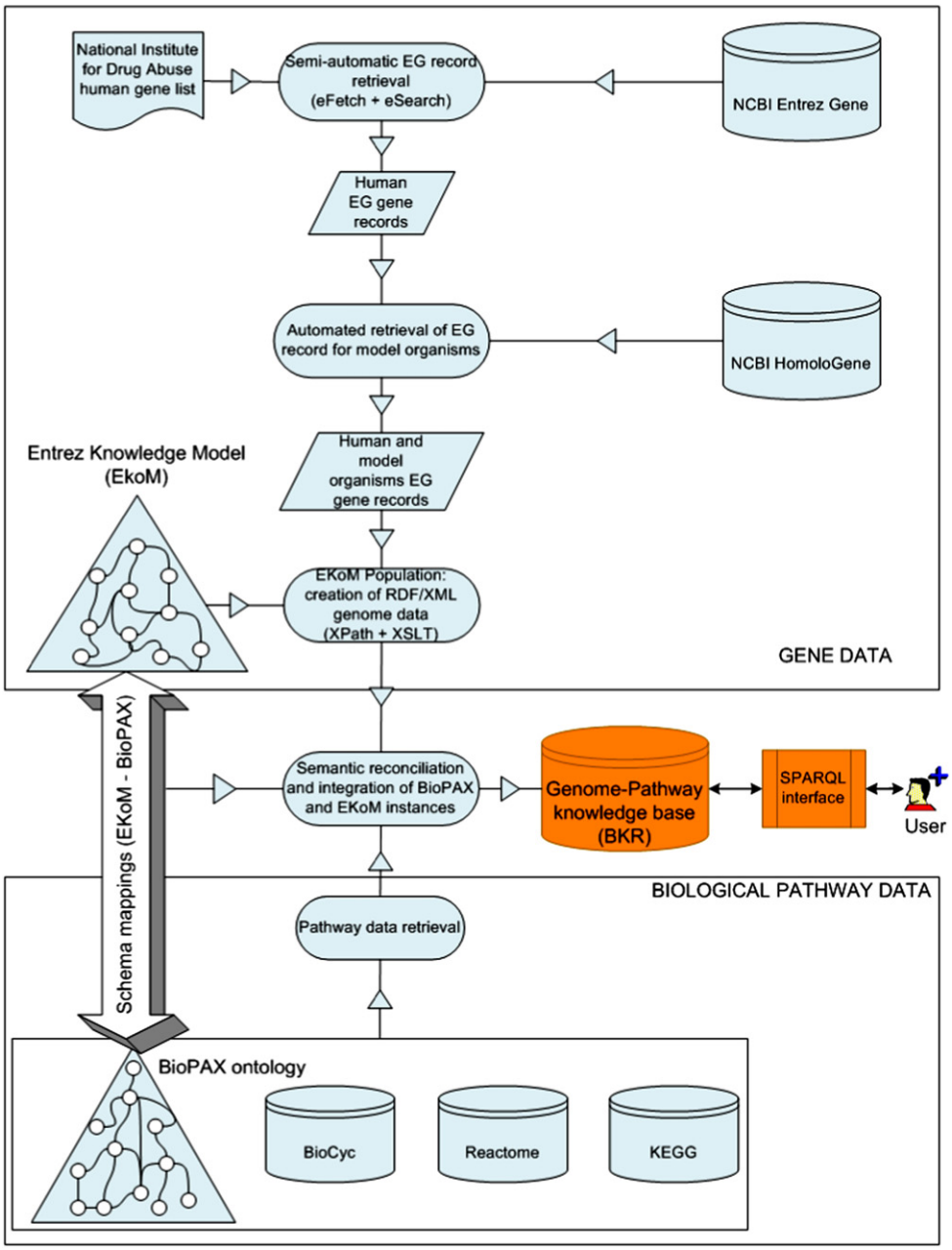

Fig. 5. Overview of the creation process for the gene pathway knowledge base.

4502827, where GeneID: 1137 is an instance of the class '*gene*' and GI: 4502827 and instance of the class '*protein*.' The process of populating the ontology is applied automatically using XPath to all the gene records. Since, the ontology population procedure is underpinned by EKoM, the consistency of the resulting knowledge base is ensured. The set of triples resulting from the conversion constitutes an RDF graph. The EKoM instance base is a 'grounded' RDF graph [29] as no blank (anonymous) nodes are created during the ontology population process.

4.4. Acquiring pathway information

The sources of pathway information used in this study include KEGG (Kyoto Encyclopedia for Genes and Genomics) [8], Reactome [47] and BioCyc [48]. As mentioned earlier, these three resources share a common information model, the BioPAX ontology, and are available in RDF, conforming to either version 1 or 2 of the BioPAX ontology.

KEGG is a large resource created by the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo. KEGG contains information for various model organisms about molecular interactions, reaction networks, cellular processes and human diseases. We restricted the extraction of KEGG data to the five organisms under investigation. KEGG is available in BioPAX level 1 format.

Reactome is a curated knowledge base of biological pathways resulting from collaboration among Cold Spring Harbor Laboratory, the European Bioinformatics Institute, and the Gene Ontology Consortium. Reactome contains various types of pathways, including metabolic, signaling, replication and regulation processes. Human pathway information in Reactome is manually curated, whereas non-human pathway information is generated by electronic pro-

jection and ortholog mapping. Importantly from an integration perspective, Reactome contains cross-references to other biological resources, including KEGG, UniProt and Entrez Gene. We restricted the extraction of Reactome data to four files corresponding to humans, *C. elegans*, mouse and fruit fly. Reactome is available in BioPAX level 2 format.

BioCyc is a collection of pathway/genome databases of predicted and curated metabolic pathways for many organisms created by the SRI Bioinformatics Research Group. The data in BioCyc is categorized based on the level of curation, namely Tier 1 (with at least one year of literature-based human curation), Tier 2 (with less than one year of literature-based curation) and Tier 3 (predicted pathways that have not undergone any curation). Two files, corresponding to humans (HumanCyc, from Tier 2) and fruit fly (from Tier 3), were integrated into the knowledge base. (The files for other organisms were not available from BioCyc site at this time). BioCyc data is available in BioPAX level 1 format.

As noted above, the three resources use different version of the BioPAX format. KEGG and BioCyc use level 1, while Reactome uses level 2. In order to ensure syntactic operability between the three pathway resources, we followed the approach suggested in [43] and converted all BioPAX namespace references in KEGG and BioCyc to the BioPAX level 2 namespace.

4.5. Implementation

The Oracle 10 g database management system provides native support for RDF and was used as our RDF store. Oracle also provides support for querying RDF triples, rule indices and indexing options for optimization. The total number of RDF triples generated in the knowledge base is about 1.5 million, with the 334,438 triples from Entrez Gene; 695,301 triples from Reactome; 175,160 triples from BioCyc and 352,793 triples from KEGG. The instance files were converted to N3 format (using the Jena API [44]) prior to being loaded into the Oracle store. The total time taken for this process was about one hour. Three indexes were created, one for each of the three components of an RDF triple: 'subject,' 'object,' and 'predicate.' Using the Oracle implementation of the SPARQL query language [18], a set of queries was executed against the knowledge base. The average query time was about 30 s with the indexes in place.

5. Queries and results

The integrated resource we created from gene and pathway information can be seen as a large graph in which the nodes are instances of the classes in the BioPAX and EKoM ontologies (e.g., genes, proteins, pathways, model organisms) and the edges are semantic relationships among these instances. This graph can be queried with query languages such as SPARQL. In practice, a SPARQL query is the formal representation of constraints on the graph. Evaluating the query against the RDF graph consists in the identification of the patterns in the graph that satisfy the query.

This work was motivated by the following three complex biological queries regarding the 449 genes putatively involved with nicotine dependence: Which genes participate in a large number of pathways? Which genes (or gene products) interact with each other? Which genes are expressed in the brain? In order to answer these questions, we created SPARQL queries, which we executed against the integrated gene-pathway resource in RDF. In this section, we present the rationale for these queries, the approach we used, and the results we obtained.

5.1. Which genes participate in a large number of pathways?

5.1.1. Rationale

This query seeks to identify hub genes, that is, those genes involved in a large number of pathways. These genes are most likely to play a particularly important role in biological systems. Downstream effectors, or proteins in the pathway, are also important as they represent the part of a particular pathway that is "closer" to the phenotypic effect of that pathway. Therefore, the objectives of this query are to (1) confirm the existence in the knowledge base of known hub genes, (2) identify proteins with which those hub genes' products interact or affect, both immediate and downstream, and (3) identify new candidates for further experimental studies.

5.1.2. Approach

For the 449 genes from the original set and their homologs in four model organisms, the RDF graph is explored to find all pathway entities from the three pathway resources. The list of genes is then ranked by the number of pathways in which the genes are involved.

5.1.3. Results

Table 1 lists the top 10 hub genes in the five model organisms under investigation. MAPK1 and MAPK3 are involved in as many as 30 pathways in *H. sapiens*, including Axon guidance and Glioma. Homology information is not always available. N/A indicates that no homologous gene is present in HomoloGene for a given gene. In many cases, no pathway information is recorded for zebrafish, despite the presence of homology information (indicated by 0).

In most cases, the hub genes in humans (*H. sapiens*) are the same as hub genes in mouse (*M. musculus*) and, to a lesser extent, in the other model organisms. However, pathways specific to a particular organism may reveal a feature specific to this organism or indicate a gap in the knowledge of other organisms. For example, although most of the pathways for CALM3 are common to human and mouse, one of them (Reactome Event: Metabolism of carbohydrates, identified by 71387) is specific to human. When no pathway information is available for a given gene in an organ-

Table 1
List of 10 genes participating in the largest number of pathways (ordered by genes in humans)

Gene Symbol (*Homo sapiens*)	Number of pathways per organism (Entrez Gene identifier)				
	Homo sapiens	*Mus musculus*	*Caenorhabditis elegans*	*Drosophila melanogaster*	*Danio rerio*
MAPK3	30 (EG:5595)	29 (EG:26417)	N/A	N/A	0 (EG:399480)
MAPK1	30 (EG:5594)	29 (EG:26413)	0 (EG:175545)	3 (EG:3354888)	0 (EG:360144)
MAP2K1	24 (EG:5604)	24 (EG:26395)	0 (EG:171872)	2 (EG:31872)	0 (EG:406728)
ALDH1A3	18 (EG:220)	5 (EG:56847)	N/A	N/A	0 (EG:751785)
ALDH2	16 (EG:217)	15 (EG:11669)	15 (EG:175691)	15 (EG:34256)	15 (EG:393462)
NFKB1	12 (EG:4790)	10 (EG:18033)	N/A	N/A	N/A
CREBBP	12 (EG:1387)	12 (EG:12914)	N/A	N/A	0 (EG:566841), 0 (EG:100002394)
CALM3	10 (EG:808)	9 (EG:12315)	N/A	N/A	0 (EG:327379)
CALM2	9 (EG:805)	9 (EG:12314)	N/A	N/A	0 (EG:368217)
CALM1	9 (EG:801)	9 (EG:12313)	N/A	N/A	1 (EG:321808)

ism, it is not possible to distinguish between the absence of study for this pathway in a given organism and a negative finding (i.e., the absence of participation of this particular gene in the pathway).

Using the information returned by this query, we created a gene-pathway network for human genes, involving 247 genes from the original set and 112 pathways (Fig. 6). Four of the genes MAPK1, MAPK3, MAP2K1, and CREBBP, involved in many pathways, are found at the center of a cluster of pathways, which provides a graphical rendering of the information in Table 1. The same view also clearly shows those pathways in which many genes from the original set participate (e.g., SNARE interactions in vesicular transport and Calcium signaling pathway).

5.2. Which genes (or gene products) interact with each other?

5.2.1. Rationale

This query also seeks to identify "hub genes", but from the perspective of gene interaction. These genes might play a particularly important role in nicotine dependence, especially if the genes with which they interact also belong to the original set. This query forms the basis for establishing networks of interacting genes.

5.2.2. Approach

Interactions among genes from various knowledge bases (HPRD, BIND, BioGrid, etc.) are recorded in Entrez Gene. A given interaction between two genes (as represented by an interaction of their gene products, or proteins) is often reported multiple times (e.g., in different sources or with different supporting evidence in the same source). For the 449 genes form the original set and their homologs in four model organisms, the RDF graph is explored to find all interactions (between one gene from the original set and another gene), with the number of mentions for each interaction. The list of genes is then ranked by the number of mentions. This query takes advantage of the modeling characteristics presented in Section 3.2 (Entrez knowledge model). More specifically, the query uses the relationship defined between two genes that are related to another gene through the property '*have_common_pathway*.'

5.2.3. Results

Five genes from the original set (CALM1, HSP90AA1, GRIN1, SNAP25 and STX1A) interact with more than ten other genes each. Fig. 7 shows an interaction network derived from the results of this query for the 449 human genes in the original set. The five top hub genes are highlighted in the network. Table 2 lists the top six interactions with the highest number of mentions. Of note, Table 2 includes mentions of interactions that are not explicitly listed in the gene records from Entrez Gene, but rather revealed from the integrated knowledge base through the harmonization (reconciliation) of interaction identifiers across sources. For example, the gene SNAP 25 [*H. sapiens*] (geneID: 6616) has fifty reported interactions in the Entrez Gene record. Forty seven additional mentions (but no additional interactions) are found for this gene in our integrated resource. Fig. 8 shows all interactions for the gene SNAP25, along with the number of mentions for each interaction.

5.3. Which genes are expressed in the brain?

5.3.1. Rationale

The neurobiology of nicotine dependence has already shown strong connections with various neurotransmitters in the central nervous system. Therefore, we want to identify those genes from the original set that are known to be expressed in various parts of the brain in order, for example, to focus subsequent experimental studies that could then examine gene function.

5.3.2. Approach

Although specific tissues may be mentioned in textual descriptions in the Entrez Gene record and would be amenable to text mining techniques, no explicit links to tissues can be easily and reliably processed. In contrast, the BioPAX ontology models the anatomical or tissue location using classes such as '*bioSource*' and

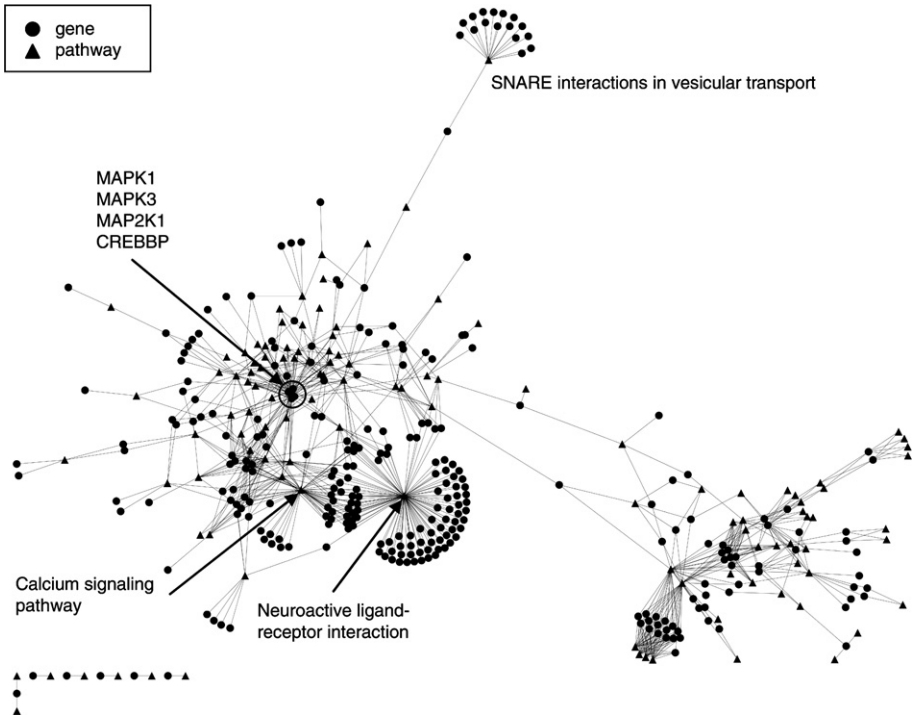

Fig. 6. Gene-pathway network for the genes from the original set.

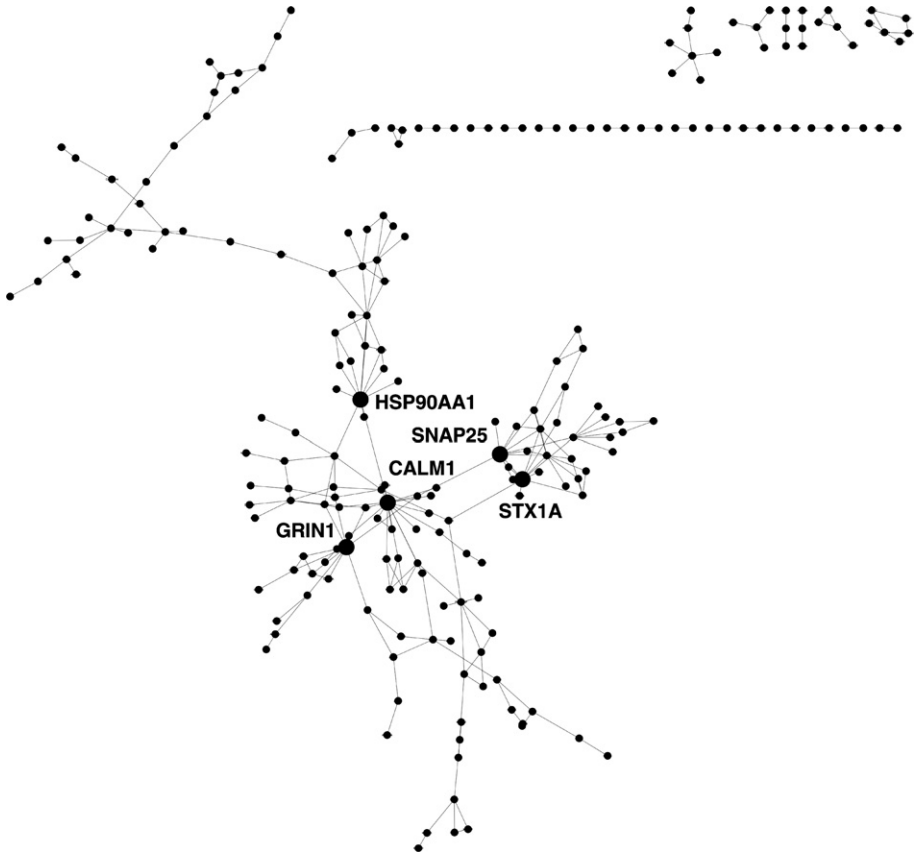

Fig. 7. Interaction network among the genes putatively involved with nicotine dependence (hub genes are highlighted and labeled).

properties such as '*TISSUE,*' both linked to *'protein.'* Starting from a gene from Entrez Gene, we can thus follow its links to proteins ('*gene*' → '*has_product*' → '*protein*'). As mentioned in Section 3.3.1, *'protein'* is common to EKoM and BioPAX and bridges between gene resources (our starting point) and pathway resources (where we find the information about tissues). However, although the concept *'protein'* is shared by EKoM and BioPAX (at the schema level), protein instances from Entrez Gene and the pathway resources use distinct identification schemes (URIs). As a consequence, it is not possible to automatically exploit tissue information in relation to genes.

5.3.3. Results

Because of heterogeneity in the identification of protein instances, our query did not return any results. However, we verified that if protein instances were reconciled (e.g., through the use of a protein-centric integrative resource such as UniProt), we would be able to link genes to tissues. For example, in Reactome, the protein Catechol *O*-methyltransferase instance is represented as *'UniProt_P21964_Catechol_O_methyltransferase__EC_2_1_1_6_.'* On the other hand, the Entrez Gene record identified the protein through its name '*catechol-O-methyltransferase.*' We manually created a mapping between the two instances (as we did to reconcile pathway instances), which enabled the traversal of the RDF graph from '*gene_1312*' (COMT Catechol *O*-methyltransferase) → '*ekom:has_product*' → *'catechol-O-methyltransferase'* → *'bp:COMMENT'* → *'...TISSUE SPECIFICITY:* ***Brain****, liver, placenta, lymphocytes and erythrocytes...'* Here again, the comment field needs to be parsed for keywords such as 'brain,' a feature supported by RDF. However, the semantics of this text field is explicit (*TISSUE SPECIFICITY*). The results are therefore likely to be reliable.

6. Discussion and future work

6.1. Technical significance

The ontology-driven framework for creating integrated knowledge bases outlined in this paper is flexible, sustainable and extensible. Answering complex biological questions typically requires manual work or the development of specific software. In contrast, we showed that the integrated resource we created can be used to answer various types of questions, demonstrating the flexibility of our approach. Because manual intervention is required only for the creation of the ontology and linkage between XML element tags and classes and relationships from the ontology, it is possible to process large volumes of data automatically and to update sources frequently. Our effort is therefore likely to be sustainable. Finally, additional information sources (e.g., transcriptome resources such as UniGene and proteome resources such as UniProt), can easily be added to the integrated resource by extending the ontology or integrating with other ontologies to accommodate new types of instances.

Some key issues encountered in this study are worth discussing and include the central role played by the ontology in typing instances, the inference of new knowledge (i.e., information gain) and the reconciliation of heterogeneous instances. But first, we want to emphasize the benefit of using Semantic Web technologies for data integration.

6.1.1. Semantic Web technologies vs. traditional approaches

This study showcases Semantic Web technologies, but could have been realized with traditional approaches as well (e.g., using relational databases techniques). In our experience, Semantic Web

Table 2
List of six gene-gene interactions, among the 449 genes putatively involved with nicotine dependence, with the largest number of mentions (human genes)

	Gene–gene interaction	Number of mentions
1.	**NR3C1** nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor) (GeneID: 2908)–**AR** androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) (GeneID: 367)	12
2	**SNAP25** synaptosomal-associated protein, 25 kDa (GeneID: 6616)–**STX1A** syntaxin 1A (brain) (GeneID: 6804)	11
3.	**SNAP25** synaptosomal-associated protein, 25 kDa (GeneID: 6616)–**VAMP8** vesicle-associated membrane protein 8 (endobrevin) (GeneID: 8673)	10
4.	**NR3C1** nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor) (GeneID: 2908)–**NR3C2** nuclear receptor subfamily 3, group C, member 2 (GeneID: 4306)	10
5.	**PTGES3** prostaglandin E synthase 3 (cytosolic) (GeneID: 10728)–**HSP90AA1** heat shock protein 90 kDa alpha (cytosolic), class A member 1 (GeneID: 3320)	10
6.	**NCOA1** nuclear receptor coactivator 1 (GeneID: 8648)–**AR** androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) (GeneID: 367)	10

technologies offer a simpler, adaptable and scalable approach to integrating biological data. In [49], we discussed the three main data integration approaches, namely, data warehousing, navigational integration and mediator-based integration. The data warehousing approach, such as GUS [50], involves importing and storing data locally in a common format. Navigational integration, exemplified by Entrez [51], creates cross-references that enables users to navigate across different data sources, interlinking resources without really integrating them. In mediator-based integration, such as in TAMBIS [14], queries are rewritten by the system before being executed against remote data sources. The approach proposed in this paper is based on Semantic Web technologies, but shares some of the features of the traditional approaches. It requires the various data sources to be converted into a common format, here RDF. The RDF store constitutes a large graph and links among entities are reminiscent of navigational integration. Finally, our approach relies on ontologies to support inference, which is also a feature of many mediator-based systems.

There are, however, several key differences between our approach and traditional data integration methods. The benefit of using Semantic Web technologies for biological data integration can be outlined as follows. Unlike databases, Semantic Web technologies provide built-in support for inference (e.g., subsumption reasoning), making it possible to infer new knowledge from existing knowledge sources (discussed later in this section). Ontologies are more expressive than database schemas and can be extended more easily. Moreover, unlike database schemas, ontologies provide a representation of entities and the relations among them that is independent of storage considerations. For these reasons, we believe that an RDF store organized around an ontology provides a simpler model and is easier for a biologist to conceptualize and query.

Although the query time was generally longer with our RDF store than expected with well-tuned relational databases, the advantages of Semantic Web technologies discussed above clearly outweigh the performance issues observed in our study. Moreover, the query time presented here must be understood as the lower bound of performance, because limited resources were spent on optimization in this feasibility study.

6.1.2. Typing instances

A reference model with well-defined, formal semantics is essential to the creation of an effective knowledge repository. XML only provides data types such as 'string', which cannot be used for reasoning purposes. It is virtually impossible, for example, to extract all instances of proteins from the Entrez Gene XML file. Therefore, in the absence of an ontology, only particular instances can be queried, not classes of instances. In contrast, typing the instances from information sources with concepts from an ontology through class-membership relations makes it possible to easily query all instances of a given class. Query 1 presented in Section 5.1 (Which genes participate in a large number of pathways?) can be used to illustrate this feature. In this query, we need to traverse the graph to find what genes are related to pathways through the relationship '*functionally_related_to*.' In the absence of an ontology to type gene instances, the entry point to the graph would necessarily be an individual gene and 449 such queries would need to be issued to find all gene-pathway associations for the 449 genes from the original set. In contrast, with all instances of genes typed with the class *'gene'* from the ontology and all instances of pathways typed with the class *'pathway'* from the ontology, this biological question requires only one query to extract all relations between instances of the class *'gene'* and instances of the class *'pathway'* through the relationship '*functionally_related_to*.' Moreover, it does not matter whether or not the '*functionally_related_to*' is used elsewhere in the ontology, as queries can put constraints on its domain and range.

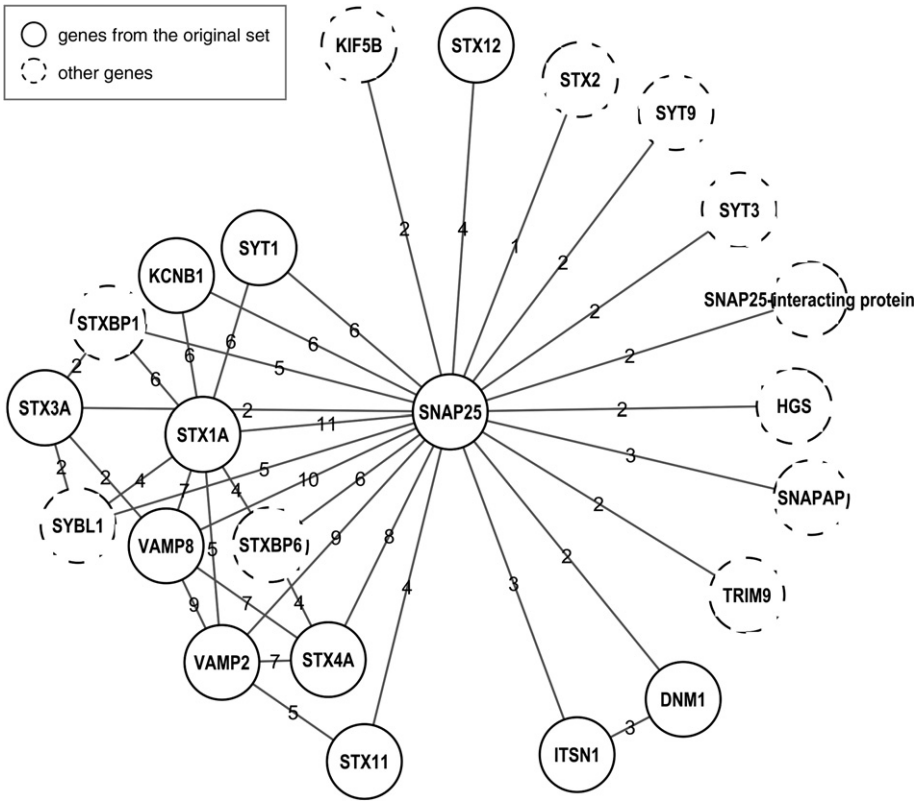

Fig. 8. Interaction network for the genes interacting with SNAP25 (the labels on the edges represent the number of mentions for each interaction).

6.1.3. Inferring new knowledge

One significant advantage of typing instances with classes from an OWL DL-based ontology is the ability to infer new knowledge from the gene-pathway knowledge base. This feature can be illustrated by the results of Query 2 presented in Section 5.2 (Which genes (or gene products) interact with each other?). Although, we did not discover any new interactions, we found additional mentions for some interactions, not recorded in Entrez Gene. In fact, additional mentions for existing interaction are important, as they increase our confidence in the existence of these interactions. These additional mentions were inferred by transitivity (i.e., if interaction identifier I identifies the interaction of gene A with gene B and the same interaction identifier I also identifies the interaction of gene B with gene C; then it can be inferred that gene A and gene C also interact). Information gain through entailment reasoning is an important advantage of ontology-based data integration. As shown in the example above, information gain can be implemented through rules on the knowledge base. The new information inferred from these rules is added to the knowledge base and extends it. Although no new interactions were discovered here, the example above illustrates the potential of ontology-based inference from biological information sources.

6.1.4. Reconciling heterogeneous instances

One important issue encountered in this study and, more generally, inherent to Semantic Web approaches to integrating resources, is the absence of a central authority or universal framework for identifying and reconciling instances. For example, the pyruvate metabolism pathway is identified by 00620 in KEGG and 71406 in Reactome. When both instances are present in an Entrez Gene record (e.g., GeneID: 4191), nothing indicates they both refer to the same pathway entity. A knowledge base created from Entrez Gene is therefore likely to contain heterogeneous instances, limiting the quality of integration.

One solution to this problem would be for the community to create a resolution service for instances, which could take the form of a common registry. This approach would be costly and difficult to maintain as the resources to be integrated evolve. Alternatively, reconciliation can be implemented locally and automatically if rules can be created from the information available in the sources. We used the latter approach in this study. As mentioned in Section 3.3.2, instances referring to the same entity in different sources were identified automatically by leveraging cross-reference information and linked together using the '*owl:sameAs*'property.

6.2. Significance for biologists

Today, a major contribution of information technologies to biology remains facilitating the work of biologists by enabling them to integrate information from heterogeneous sources and process large amounts of data. This integration then can facilitate the generation of new hypotheses about biological significance to be confirmed by future experiments.

We showed that it was possible to integrate gene and pathway information into a common framework and to reason over the integrated resources, taking advantage of an ontology to ensure the consistency of and facilitate queries against the knowledge base we created. We also showed that standard tools and technologies could support various types of queries and that the information returned by these queries could be easily converted to produce the kinds of representations used by biologists (e.g., interaction networks created with Cytoscape).

Among the 449 genes from the original set, only 247 are linked to pathways described in the three sources integrated in our knowledge base (Fig. 6). Analogously, gene–gene interactions are

recorded in our resources for only 219 genes from the original set (Fig. 7). In addition to hub genes, likely to play a particularly important role in biological systems, the genes identified by genome-wide linkage and association studies as potentially related to nicotine dependence for which no interactions to other genes and links to pathways are currently recorded probably deserve the attention of researchers.

This study exploits existing knowledge and was not expected to result in any new findings. However, our findings corroborate analyses of some of the same sources (e.g., KEGG) independently conducted by other researchers using different techniques. For example, the pathways Neuroactive ligand–receptor interaction and Calcium signaling pathway highlighted in Fig. 6 are also mentioned by [52] as being significantly enriched for addiction-related genes. The genes from the MAPK group and the CREBBP gene, which form a cluster of hub genes in Fig. 6 are also cited by [52]. Analogously, the five pathways listed by [52] in relation with nicotine dependence (Neuroactive ligand–receptor interaction, Long-term potentiation, GnRH signaling pathway, MAPK signaling pathway and Gap junction) are all present in our dataset and all appear to be linked to at least 15 genes from our original dataset.

6.3. Limitations and future work

The study presented here has several limitations, which we plan to address in the future, regarding the heterogeneity of instances, the identification of anatomical information and the absence of integration between structured information sources and the biomedical literature.

6.3.1. Heterogeneity of instances

As mentioned earlier, the presence in the knowledge base of distinct instances referring to the same entity results in limited integration between information sources. In this study, we observed this phenomenon mostly for proteins and pathways. Proteins were generally identified by their name, making it difficult to match them exactly and reliably across resources. Pathways, on the other hand, were generally identified with identifiers local to a given resource, making it impossible to relate them across sources in the absence of a mapping service. We solved the problem in part for pathways by exploiting the cross-reference information provided in some sources, such as Reactome. We would need to integrate additional information sources to bridge across namespaces for proteins. We plan to integrate UniProt for this purpose.

6.3.2. Anatomical information

The reason why Query 3 (Which genes are expressed in the brain?) was not successful is not because of the absence of anatomical information, but rather because it was extremely difficult to bridge between proteins across resources due to instance heterogeneity. Moreover, extracting anatomical information from the comments field in Reactome, related to the '*protein*' concept in pathway resources by '*bp:COMMENT*' relationship, was possible, but not straightforward, as the field had to be parsed for keywords. However, had proteins been perfectly integrated and anatomical information been present in a specific string, queries would still have been suboptimal, due to the absence of a reference ontology of anatomy. In fact, to a biologist, the query "expressed in the brain" is actually a shortcut for "expressed in the brain or any of its parts." An ontology of anatomy such as the Foundational Model of Anatomy [53] would support the expansion of the query by exploiting subclass and partonomic relations among anatomical entities. More generally, reasoning over specialized information sources such as gene and pathway resources often benefits from reference ontologies for domains such as anatomy, diseases, and drugs.

6.3.3. Integrating knowledge extracted from the biomedical literature and structured knowledge bases

This work is a pilot contribution to the Biomedical Knowledge Repository under development at the U.S National Library of Medicine (NLM) as part of the Advanced Library Services project [54]. This repository integrates knowledge not only from structured resources (database and knowledge bases), but also from the biomedical literature (e.g., MEDLINE), in order to support applications, including knowledge discovery. This study is limited to the information extracted from five structured information sources. However, we are working on the integration of knowledge extracted from MEDLINE citations. To select the appropriate corpus from MEDLINE, we plan to use not only a PubMed search on "nicotine dependence," but also the list of PubMed identifiers (PMIDs) cited as evidence by the curators of the pathway information sources. Combining these two sources of information, structured and unstructured, is expected to fill the gaps observed in pathway resources for some organisms.

7. Conclusion

Semantic Web technologies provide a valid framework for information integration in the life sciences. We illustrated how two gene information sources (Entrez Gene and HomoloGene) and three pathway information sources (KEGG, Reactome and BioCyc) can be integrated into a knowledge base using RDF for its representation. Ontology-driven semantic integration represents a flexible, sustainable and extensible solution to the integration of large volumes of information. Because instance entities are typed with classes from the ontology, ontology-driven integration ensures the consistency of the knowledge base and facilitates the query process. For example, we showed that queries could be formulated for the class '*gene*' as a whole, not only for individual gene instances. This work also illustrates the versatility of the integration framework, as no specific tools are required to produce results that can be imported in the tools used by biologists for visualization purposes. The limitations encountered in this study can be compensated for by integrating additional resources to bridge across namespaces (e.g., UniProt), to support reasoning (e.g., anatomical ontologies), and to broaden the scope of information sources (e.g., with information extracted from the biomedical literature).

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). The authors want to thank Lee Peters who created the RDF store in Oracle and helped craft and run the SPARQL queries and Jonathan Pollock for valuable input on model organism databases. Cytoscape was used to visualize the interaction networks.

References

[1] Hall W, Madden P, Lynskey M. The genetics of tobacco use: methods, findings and policy implications. Tob Control 2002;11(2):119–24.
[2] Li MD, Ma JZ, Cheng R, Dupont RT, Williams NJ, Crews KM, et al. A genome-wide scan to identify loci for smoking rate in the Framingham Heart Study population. BMC Genet 2003;4(Suppl. 1):S103.
[3] Tyndale RF. Genetics of alcohol and tobacco use in humans. Ann Med 2003;35(2):94–121.
[4] Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet 2007;16(1):24–35.

[5] Li MD. The genetics of nicotine dependence. Curr Psychiatry Rep 2006;8(2):158–64.
[6] Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet 2007;16(1):36–49.
[7] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005;33(Database Issue):D54–8.
[8] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resources for deciphering the genome. Nucleic Acids Res 2004;32:D277–80.
[9] http://www.ncbi.nlm.nih.gov/sites/entrez/query.fcgi?db=homologene. 22 January 2008.
[10] Hey T, Trefethen AE. Cyberinfrastructure for e-Science. Science 2005;308(5723):817–21.
[11] Luciano J. PAX of mind for pathway researchers. Drug Discov Today 2005;10(13):937–42.
[12] Cheung K-H, Smith AK, Yip KYL, Baker CJO, Gerstein MB. Semantic Web approach to database integration in the life sciences. In: Baker CJO, Cheung K-H, editors. Semantic Web: revolutionizing knowledge discovery in the life sciences. New York: Springer; 2007. p. 11–30.
[13] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung K-H. Advancing translational research with the Semantic Web. BMC Bioinformatics 2007;8(Suppl. 3):S2. PMID:17493285.
[14] Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 2000;16(2):184–5.
[15] Zhao J, Goble C, Stevens R. Semantic web applications to e-science in silico experiments. In: Proceedings of the 13th international world wide web conference on alternate track papers & posters 2004, May 19–21, 2004; New York, NY, USA: ACM Press; 2004. p. 284–5.
[16] http://www.w3.org/TR/owl-features/. 22 January 2008.
[17] http://www.w3.org/TR/rdf-primer/. 22 January 2008.
[18] Prud'ommeaux E, Seaborne, A. SPARQL Query Language for RDF. World Wide Web Consortium 2004.
[19] Sirin E, Parsia B, Cuenca Grau B, Kalyanpur A, Katz Y. Pellet: a practical OWL-DL reasoner. J Web Semant 2007;2(5).
[20] Haarslev V, Möller, R. Racer: a core inference engine for the Semantic Web. In: Proceedings of 2nd international workshop on evaluation of ontology-based tools (EON2003), located at the 2nd international semantic web conference ISWC 2003, 2003 October 20, Sanibel Island, Florida, USA; 2003. p. 27–36.
[21] http://esw.w3.org/topic/HCLS/Banff2007Demo. 22 January 2008.
[22] Luciano JS, Stevens RD. e-Science and biological pathway semantics. BMC Bioinform 2007;8(Suppl.):S3.
[23] Bichindaritz I. Mémoire: case based reasoning meets the semantic web in biology and medicine. Heidelberg: Springer Berlin; 2004.
[24] Cheung K, Qi P, Tuck D, Krauthammer M. A semantic web approach to biological pathway data reasoning and integration. Web Semant 2006;4(3):207–15.
[25] Wolstencroft KJ, Stevens R, Tabernero L, Brass A. PhosphaBase: an ontology driven database resource for protein phosphatases. Proteins 2005;2(58): 290–4.
[26] Aleman-Meza B, Burns P, Eavenson M, Palaniswami D, Sheth AP. An ontological approach to the document access problem of insider threat. In: IEEE international conference on intelligence and security informatics (ISI-2005); 2005.
[27] Perry M, Hakimpour F, Sheth A. Analyzing theme, space and time: an ontology-based approach. In: Proceedings of fourteenth international symposium on advances in geographic information systems (ACM-GIS '06); 2006 November 10–11; Arlington, VA; 2006.
[28] Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. TAMBIS—transparent access to multiple bioinformatics information sources. Intell Syst Mol Biol 1998(6):25–34.
[29] http://www.w3.org/TR/rdf-mt/#defentail. 22 January 2008.
[30] Lenzerini M. Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGART symposium on principles of database systems; 2002 June 3–5; Madison Wisconsin: ACM Press; 2002. p. 233–46.
[31] Sahoo SS, Zeng K, Bodenreider O, Sheth AP. From "glycosyltransferase" to "congenital muscular dystrophy": integrating knowledge from NCBI Entrez Gene and the Gene Ontology. In: KKea, editor. Medinfo; 2007; Brisbane, Australia: IOS Press; 2007. p. 1260–4.
[32] http://www.w3.org/TR/xpath. 22 January 2008.
[33] http://www.w3.org/TR/xslt. 22 January 2008.
[34] Goble C, Wolstencroft K, Goderis A, Hull D, Zhao J, Alper P, et al. Knowledge discovery for biology with Taverna: producing and consuming semantics in the Web of Science. In: Baker CJO, Cheung K-H, editors. Semantic web: revolutionizing knowledge discovery in the life sciences. New York: Springer; 2007. p. 355–95.
[35] Stein L. Creating a bioinformatics nation. Nature 2002;417:119–20.
[36] Wilkinson MD, Links M. BioMOBY: an open-source biological web services proposal. Brief Bioinform 2002;3(4):331–41. PMID: 125110062.
[37] Chabalier J. Mosser J, Burgun A. Integrating biological pathways in disease ontologies. In: Medinfo, 2007. IOS Press; 2007. p. 791–5.
[38] Gudivada RC, Qu XA, Jegga AG, Neumann EK, Aronow BJ. A genome–phenome integrated approach for mining disease-causal genes using semantic web. In: HCLS Workshop, WWW 2007. Banff Canada: ACM Press; 2007.
[39] Ruttenberg AR, Rees JR, Luciano JS. Experience Using OWL DL for the exchange of biological pathway information. In: Workshop 2005 OWL: experiences and directions. Galway, Ireland; 2005.
[40] http://www.biopax.org/. 22 January 2008.
[41] Mitchell JA, McCray AT, Bodenreider O. From phenotype to genotype: issues in navigating the available information resources. Methods Inf Med 2003;42(5):557–63.
[42] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol 2005;6(5):R46.
[43] Kotecha N, Bruck K, Lu W, Shah N. Pathway knowledge base: integrating BioPAX compliant data sources. In: HCLS Workshop, ISWC 2006, Athens, GA; 2006.
[44] McBride B. Jena: a semantic web toolkit. IEEE Internet Comput 2002;6:55–9.
[45] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html. 22 January 2008.
[46] Sahoo SS, Bodenreider O, Zeng K, Sheth AP. Adapting resources to the semantic web: experience with Entrez Gene. In: Workshop on semantic web health care & life sciences at ISWC 2006, Athens, GA, USA; 2006.
[47] Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol 2007;8(R39).
[48] Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 2004;6(R2):1–17.
[49] Sahoo SS, Bodenreider O, Zeng K, Sheth AP. An experiment in integrating large biomedical knowledge resources with RDF: application to associating genotype and phenotype information. In: Workshop on health care and life sciences data integration for the semantic web at the 16th international world wide web conference (WWW, 2007), Banff, Canada; 2007.
[50] Davidson SB, Crabtree J, Brunk BP, Schug J, Tannen V, Overton GC, et al. K2/Kleisli and GUS: experiments in integrated access to genomic data sources. IBM Syst J 2001;40(2):512–31.
[51] Entrez. In.
[52] Li C, Mao X, Wei L. Genes and (Common) Pathways Underlying Drug Addiction. PLoS Comput Biol 2008;4(1).
[53] Rosse C, Mejina Jr JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(2003):478–500.
[54] Bodenreider O, Rindflesch TC. Advanced library services: developing a biomedical knowledge repository to support advanced information management applications. Technical report. Bethesda, Maryland: Lister Hill National Center for Biomedical Communications, National Library of Medicine; 2006 September 14.