DRUG DISCOVERY TODAY BIOSILICO

PAX of mind for pathway researchers

Joanne S. Luciano

Scientists seeking to understand the inner workings of cells have access to a multitude of pathway data resources. However, the representations of pathway data within these resources are not consistent or interchangeable. To facilitate easy information retrieval from a wide variety of pathway resources, such as signal transduction, gene regulation, molecular interaction and metabolic pathway databases, a broad effort in the biopathways community called BioPAX was formed. New biological pathway software applications built using the BioPAX standard will be able to integrate knowledge from multiple sources in a coherent and reliable way. This article reports the progress that the BioPAX work-group has made towards building and deploying the BioPAX data-exchange format for biological pathway data.

Databases containing information about biological pathways are almost as many and varied as the pathways themselves [1]. Even the term 'pathway' has multiple interpretations: some consider a pathway to be a network of interactions; others see a pathway as a chain of biochemical reactions linked together by substrates and products. Regardless of the definition used, access to a single integrated source of pathway data could streamline the work of scientists who apply pathway information in their work.

But how can such a resource be built when there are different concepts of what a pathway is? The four main categories of pathway data today are metabolic pathways, molecular interactions, gene regulation networks and signaling pathways, which are all traditionally represented differently. Metabolic pathways are usually shown as a series of enzyme-substrate-product reactions; molecular interactions, such as protein–protein interactions obtained from yeast two-hybrid (Y2H) experiments, are usually depicted as simple binary interactions; gene regulation pathways show connections between transcription factors and the genes whose transcription they activate or repress; signaling pathway representations are the most varied, ranging from vague and general representations of the form 'there's an activation chain in which A activates B activates C' to specific and detailed representations involving a series of complex binding reactions and protein post-translational modifications.

In order to combine different concepts of biological pathways into one consistent specification, or ontology, the representation must be detailed enough to express subtle properties of pathways, yet general enough to maintain the framework when few details are understood.

Pathway data silos

Many databases of biological pathway data exist and the number is growing. The Pathway Resource List (http://cbio.mskcc.org/prl), as of April 2005, referenced 176 pathway databases, up from 138 in July 2004. However, the number of databases is not the problem; the problem is that each one uses its own data model (semantics) and data format (syntax) for the data it provides, and each needs to be individually

Joanne S. Luciano, PhD
Harvard Medical School,
BioPathways Consortium,
BioPAX Workgroup,
Predictive Medicine, Inc.
45 Orchard Street,
Belmont, MA 02478, USA
e-mail:
jluciano@biopathways.org

1359-6446/05/$ – see front matter PII: S1359-6446(05)03501-4

translated and merged with other data sets to be suitable for large-scale analysis.

For instance, consider that a drug company might be interested in studying potential drug targets from a particular metabolic pathway, including the enzymes that catalyze the reactions, the transcription factors that activate the enzyme genes, the signal transduction pathway proteins that activate the transcription factors and the cell-surface receptor proteins that initiate signaling. The problem is that each of these drug targets may reside in a different pathway database and must be translated from its native format into a common format that enables integration. This format must be capable of representing each type of drug target and the biological context in which it exists. It must also be capable of resolving the original source from which each drug target comes, while preserving the identity of the target, regardless of its multiple names.

These are the challenges that BioPAX seeks to address. BioPAX (a creative acronym for *Bio*logical *PA*thway data-e*X*change format) is a specification for representing signal transduction, gene regulation, molecular interaction and metabolic pathways to enable coherent and reliable queries across multiple databases.

The BioPAX definition of a pathway is deliberately generic, defining a pathway as a set or series of interactions or reactions, often forming a network, which biologists have found useful to group together. This definition captures the essential characteristics common to the many different pathway representations – that of a biologically meaningful collection of interactions. Rather than placing the burden of integration upon the data consumers (i.e. drug companies, research laboratories, software vendors), the goal of BioPAX is to deal with the problem at its source by helping data providers export their data in a common format.

Another goal of BioPAX is to enable distributed curation. For example, if curators know that the data they are generating can and will be shared, and that they will have access to other curated data, it becomes possible and advantageous for all parties to share the workload and eliminate the duplication of effort in each curating the same pathway.

BioPAX development and implementation

A few conditions are crucial to developing a new data standard: a small group of dedicated individuals, commitment from the major data providers, financial support and community buy-in. The idea for the BioPAX initiative can be traced back to a speech by Chris Sander at the Intelligent Systems for Molecular Biology (ISMB) meeting in 2001 (http://ismb01.cbs.dtu.dk/), whose keynote address highlighted the need for a public repository of pathway information. It wasn't until a year later, however, at ISMB 2002, that interested individuals, database providers and users got together and agreed to take the first step: to develop and adopt a common data-exchange format for pathway data (which was later dubbed BioPAX). Coincidentally, it was also at this meeting that BioPAX, through the BioPathways Consortium (BPC; www.biopathways.org), gained the interest and support of the US Department of Energy (DOE). Thus, BioPAX quickly achieved its initial goals of securing funding and the commitment of two major pathway databases: BioCyc, a collection of metabolic pathway databases developed by Peter Karp's group at SRI International that includes the MetaCyc [2] and EcoCyc [3] databases; and BIND [4], a molecular interactions database developed by Chris Hogue's group at the University of Toronto.

However, soliciting the knowledge and support of many other pathway data providers and consumers, ontology developers and software tool developers would be key to the success of the project. To reach the rest of the biopathways community who were not present at the initial meeting, a core group of BioPAX coordinators gave presentations and held *ad hoc* sessions at international meetings, managed the web site and mailing list, gave workshops and seminars, and engaged in other outreach activities. Hence, the early vision of a public pathway database resource evolved into a plan to first provide a data-exchange format for the biopathways community.

Subsequently, the BioPAX group developed a list of technical goals for the format. Chief among these were flexibility, so that a wide range of data could be represented; computability, so that software can read the BioPAX specification and automatically understand how to compute with BioPAX data; and compatibility with existing standards, so that BioPAX would be as easy as possible for users to adopt.

To ensure compatibility with existing pathway standards, such as the Proteomics Standards Initiative molecular interaction (PSI-MI) format [5] and Systems Biology Markup Language (SBML) [6], the BioPAX work-group recruited members of each standard to help develop BioPAX. As a result, the PSI-MI standard formed the basis of BioPAX level 2, and the reactions and species of an SBML model can now be annotated with BioPAX metadata.

The BioPAX concepts were specified in the Web Ontology Language (OWL) using two ontology editing tools: Generic Knowledgebase Editor (SRI International; http://www.ai.sri.com/~gkb/ [7]) and Protégé (Stanford Medical Informatics; http://protege.stanford.edu/) with the Protégé OWL plug-in (http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf). OWL is an extension of the Resource Description Framework (RDF), both of which are standards recommended by the World Wide Web Consortium (W3C). OWL/RDF uses a simple subject-predicate-object format to represent data about data or metadata. Using this flexible 'triple' format, it is easy to focus on the meaning of the data, rather than its format (see Box 1).

BOX 1

XML, RDF and OWL

XML: eXtensible Markup Language
In HTML, tags give meaning and structure to a web document. XML is similar to HTML, except that, in XML, you can define the tags (this is why it is called extensible). The structure of an XML document can be validated with either an XML schema document or the document type definition (DTD) specification. XML schemas and DTDs describe the XML tags and the allowable structure of an XML document (http://www.w3.org/XML/).

RDF: Resource Description Framework
RDF uses XML syntax and is formatted as an XML document (angle-bracketed tagged data); however, that is where the similarity ends. Whereas XML documents can be represented as trees, RDF documents are graphs. All data in RDF are described using subject-verb-object triples, which define the 'semantics' or 'domain logic' needed to connect various data items and specify their relationship to each other. The objects that RDF describes are called universal resource identifiers (or URIs), which resemble web addresses. This is where the power of RDF comes from. As each triple refers to a single subject-predicate-object 'fact', one can assemble all the facts into a web of information. If everyone else publishes their facts, but they reuse the same URIs, then what results is the globally distributed network called the semantic web. See http://www.w3.org/ and http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

OWL: Web Ontology Language
OWL, which is defined using RDF, is a language designed for ontology construction and deployment. OWL adds the required semantic constraints on the RDF language(s) used for data documents. Together, RDF and OWL form a logic model that can be used throughout either data repositories or knowledge bases and inference engines. The structure and meaning of any group of documents can be precisely defined and related to each other. For example, similar models, one using 'mother' and the other using 'female parent', can be semantically linked using statements such as owl:isEquivalentTo. Similarly, one could use another OWL statement, owl:disjointClass, to describe the fact that the two concepts 'mother' and 'father' were disjoint (i.e. noone can be both a mother and a father). One could also use cardinality constraints to restrict the number of mothers one may have to exactly one, or one could create subclasses of mother to be surrogate, step, biological and genetic. The advantage of OWL is that you get these rich semantics in a machine-readable format. See http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html [16]. If you are interested in learning about Protégé OWL, go to http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf [17].

Expressing metadata in RDF is easier than building an ontology in OWL. RDF, like XML, is semi-structured and may be more useful in the early stages of research, when your understanding of the restrictions on relationships between the entities isn't clear or can change rapidly. When your research has reached the stage where you can articulate your understanding and capture that knowledge, then investing the time to build a full-blown ontology in OWL will enable you to more efficiently share your knowledge and utilize reasoning to advance your knowledge even further.

The BioPAX ontology is based on several existing database schemas, most notably aMAZE [8], BIND [4], EcoCyc [3], WIT [9] (now PUMA2), KEGG [10] and Reactome [11]. The ontology supports a wide range of detail and multiple levels of abstraction, which is important because BioPAX needs to support a variety of data models. Currently, semantic mapping is done manually – the mapping of the database fields from the native database format to the BioPAX format typically requires at least one expert from the source database and at least one expert in BioPAX. The BioPAX community is developing tools to facilitate this process. One such tool under development is libBioPAX, an application program interface (API) to automate the creation of BioPAX OWL instances from source data. Ideally, all pathway databases will make their data available in BioPAX format. The BioPAX ontology can be downloaded from the web site (www.biopax.org) and is available under the GNU Lesser General Public License (http://www.gnu.org/copyleft/lesser.html).

The BioPAX deliverable was partitioned into levels of increasing complexity to be tackled sequentially. BioPAX level 1, which was released in July 2004, focused on representing metabolic pathways, because they are the oldest, most widely used and best understood type of pathway. Tackling metabolic pathways first made the largest body of data available in the shortest possible time. As specific subtopics of BioPAX development arose, the BioPAX work-group formed subgroup task forces. Each subgroup typically contained one or more BioPAX coordinators and several outside experts. The BioCyc family of databases has already been made available in BioPAX. BIND, PUMA2 (http://compbio.mcs.anl.gov/puma2), aMAZE, Reactome and KEGG are all in various stages of conversion.

Level 2, released as a beta version in April 2005, implements molecular interactions and adopts many features of the PSI-MI format (http://psidev.sourceforge.net/mi/xml/doc/user/). It initially covered only protein–protein interactions, but now includes other molecular interactions, such as protein–DNA and protein–RNA interactions. BioPAX level 3 will address signaling pathways and work-group participants from around the globe met in January 2005 to begin development (http://www.biopax.org/Docs/Jan05mtng). By developing BioPAX in levels – first metabolism, then protein interactions, followed by signal transduction and gene regulation – the standard will grow as consensus is reached on how to represent progressively more complex pathway types. The BioPAX road map (Table 1) provides details about what types of pathways are included in the ontology and what kinds of databases will be supported with each addition.

The BioPAX ontology

The top level of the BioPAX ontology is the root class, 'entity', which BioPAX defines as 'a discrete biological unit used to describe biological pathways'. All other biological

TABLE 1

BioPAX road map

Development level	Scope of format	Sample data sources[a]
Level 1	Metabolic pathways	aMAZE, BioCyc, KEGG, PUMA2
Level 2	Level 1 plus molecular interactions	BIND, DIP, HPRD, IntAct, MINT
Level 3	Level 2 plus signaling pathways and gene regulation	CSNDB, INOH, PATIKA, Reactome, TRANSPATH
Level 4	Level 3 plus genetic interactions	FlyBase, MIPS
Future levels	Level 4 plus abstract associations	PubGene, GeneWays

[a]For a complete listing, see http://www.cbio.mskcc.org/prl.

concepts in the BioPAX ontology – physical entities, interactions and pathways – are subclasses of this root class.

Physical entities are the building blocks of simple interactions. For example, small molecules, RNA and proteins are all considered physical entities. Complexes, which are composed of other physical entities, are also themselves physical entities.

In biology, many different kinds of interactions are possible and it is important to describe the details of any given interaction unambiguously. For example, with the BioPAX ontology, one can specify that a biochemical reaction is catalyzed by two isoenzymes, the first of which is a protein complex composed of multiple subunits, and that the catalysis of the second enzyme is modulated by a small molecule. This is possible because the type of physical entity (small molecule, protein, complex) is separate from the role that it plays in the interaction (complex subunit, substrate, product, catalysis modulation and biochemical reaction catalysis).

Pathways in BioPAX are entities composed of a set of interactions. The top-level BioPAX definition of pathway is general enough to capture the many kinds of pathways used by biologists.

Data as instances

Pathway data represented in BioPAX are considered instances of the BioPAX ontology. For example, the BioPAX ontology defines a biochemical reaction as something that has two important properties: a left- and a right-hand side. Thus, every unique reaction in BioPAX must have a unique set of left- and right-hand side participants (Figure 1).

BioPAX reuses existing standards wherever possible. For example, instead of building a list of possible cellular locations (nucleus, cytoplasm, etc.) directly into the BioPAX ontology, the location property in BioPAX is able to point to a controlled vocabulary term from the cellular component branch of the Gene Ontology (GO; www.geneontology.org [12]), which provides an exhaustive, expertly curated list of cellular locations.

BioPAX also makes use of the Open Biomedical Ontologies (http://obo.sourceforge.net/) cell type ontology, cell.obo, to provide a controlled vocabulary for cell types and the NCBI taxon database (ftp://ftp.ncbi.nih.gov/pub/taxonomy/) as its source for organism names. In some cases, BioPAX encapsulates other data standards, as opposed to merely referencing them. For example, in order to represent chemical structures, BioPAX permits SMILES (www.daylight.com/smiles/) strings [13], CML (www.xml-cml.org) structure definitions [14] and InChI, the new IUPAC/NIST Chemical Identifier (www.iupac.org/projects/2000/2000-025-1-800.html) [15].

In general, when vocabulary terms or structure representations are defined elsewhere and widely accepted as the standard, BioPAX adopts them. The BioPAX workgroup recognized early on that it would be best to let experts define the standard vocabularies for their area of expertise and to focus their attention on building the framework for bringing different standards together.

The semantic web and data integration

BioPAX is looking ahead to the semantic web. Because BioPAX was specified using OWL/RDF, it will enable web services to seamlessly query and join BioPAX data from multiple locations. These technologies are still in the early stages of development and may require additional commitment from data providers, such as providing stable Life Science Identifiers (LSIDs) for their data objects. If and when semantic web infrastructure is mature enough to justify these commitments, BioPAX will be well poised to become the language of choice for describing biological metadata on the semantic web.

BioPAX may also be ripe for incorporation into commercial applications. By publishing data in the BioPAX ontology, important results can be communicated to the public without exposing underlying proprietary ontologies or data models. One assumption motivating the development of BioPAX is that the competitive advantage to each corporation lies in the knowledge that is contained in the data format, not the data format itself – no one benefits from many mutually incomprehensible languages.

Finally, as integration across pathway representations grows increasingly necessary, standards for interoperability become more critical. BioPAX includes a number of constructs specifically designed to support database integration, for example, a type of external reference called a 'unification xref'. By definition, two objects that contain unification xrefs that point to the same object (e.g. a protein in the Swiss-Prot database) are themselves the same object. This feature allows users to identify

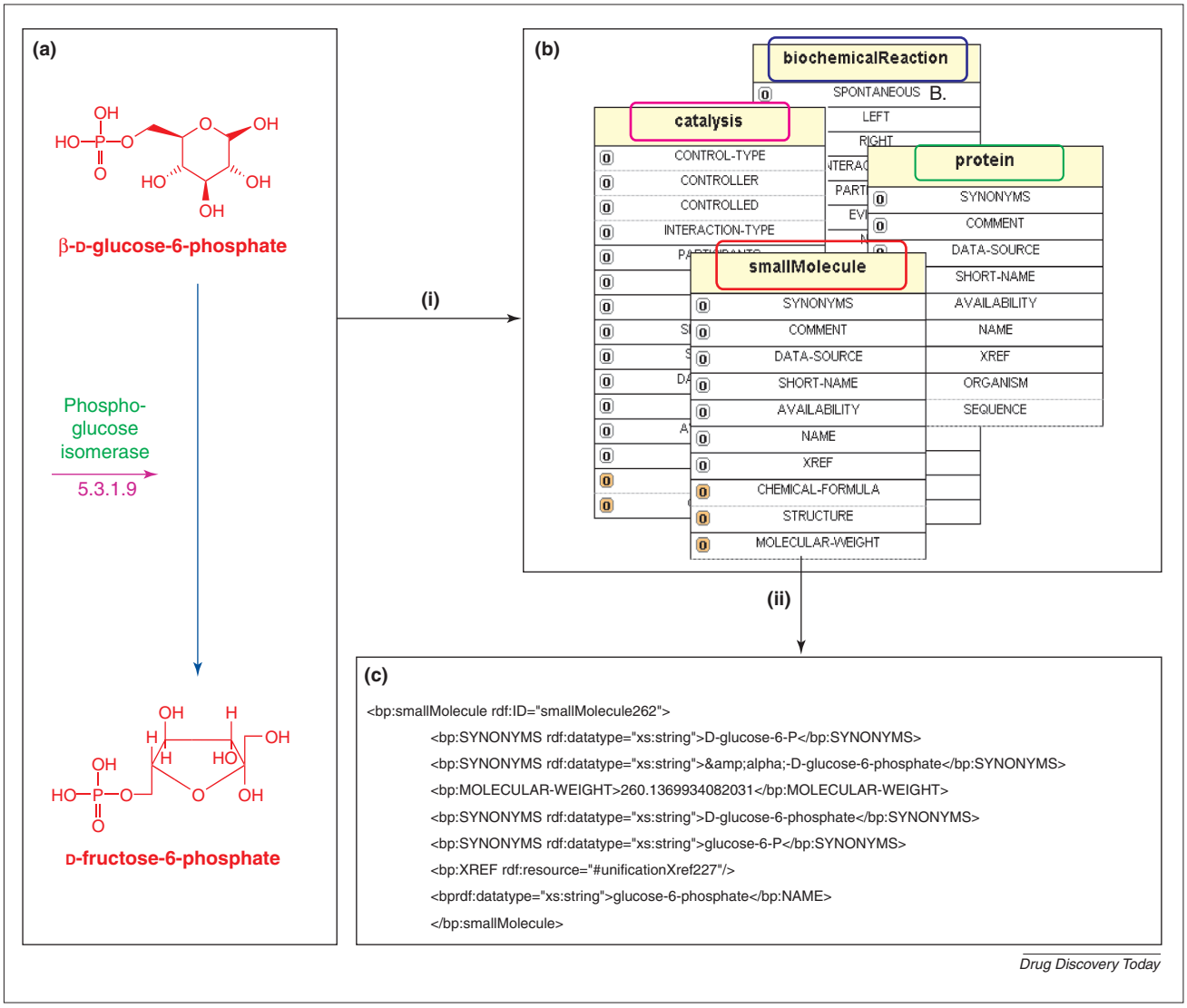

FIGURE 1

Representing pathway data in BioPAX. (i) Pathway information, such as a single step in the glycolysis pathway **(a)**, is mapped to specific classes in the BioPAX ontology **(b)**. Colors of objects in **(a)** correspond to highlighted classes in **(b)**. **(ii)** BioPAX classes serve as information templates for the generation of data instances, which are stored in OWL and represented as RDF/XML **(c)**. Pathway visualizations such as that shown in **(a)** are human readable and, in many cases, are graphical images that are not machine readable. RDF and XML, as shown in **(c)**, are the opposite. They are highly machine readable, but not very helpful to humans.

instances in different BioPAX data sets that are alternative representations of the same biological object, which is a crucial step in the process of integrating data from multiple sources.

The future of BioPAX

Although the main function of BioPAX is the exchange of data between biological pathway databases, increasing interest and investment in disciplines that rely on pathway data, such as systems biology, could lead to a wide variety of additional uses.

For example, BioPAX could be used to annotate SBML data. SBML is an XML-based data format designed to exchange pathway models between software simulation packages. Pure SBML supports only minimal descriptions of molecular species, but these descriptions could be enhanced by incorporating elements of BioPAX directly into the SBML code. By matching SBML meta-identifiers to BioPAX RDF identifiers, users may specify such things as the nature of molecules (protein, small molecule, etc.), references to external databases and name synonyms.

This notion of using one standard to extend and enhance another is the beginning of a new kind of data integration, fueled by semantic web technologies such as OWL, RDF and LSIDs (http://lsid.sourceforge.net/). As BioPAX is based on OWL, the language of the semantic web, it will probably facilitate many sophisticated applications beyond simple sharing of pathway data, such as computational reasoning on pathway data. If semantic web technologies develop as promised, BioPAX could lay the foundation for a vast network of interconnected

biological information. This network would be fully accessible and interpretable by pathway analysis software, or 'pathway agents', which could dramatically accelerate the pace of pathway research.

A community effort

BioPAX is a community effort and welcomes participation by interested groups and individuals. To participate in BioPAX workshops, technical discussions, decision making, and design and development of tools and validation software, join the biopax-discuss mailing list (http://www.biopax.org/mailman/listinfo/biopax-discuss). To keep apprised of major BioPAX developments, join the biopax-announce mailing list (http://www.biopax.org/mailman/listinfo/biopax-announce). To participate in the administration of BioPAX or to sponsor BioPAX activities, send an e-mail to biopax-dev@googlegroups.com. All BioPAX documentation, including BioPAX presentations, meeting and conference call minutes, and lists of participating individuals and organizations, can be found on the BioPAX web site (http://www.biopax.org).

Acknowledgements

BioPAX is the work of the BioPAX work-group. Major contributors to BioPAX level 1, the focus of this article, were Gary D. Bader, Erik Brauner, Michael P. Cary, Robert Goldberg, Chris Hogue, Peter Karp, Joanne Luciano, Debbie Marks, Natalia Maltsev, Eric Neumann, Suzanne Paley, John Pick, Aviv Regev, Andrey Rzhetsky, Chris Sander, Vincent Schachter, Imran Shah, Mustafa Syed and Jeremy Zucker. Special thanks to the remaining members of the BioPAX community and work-group, to the Office of Biological and Environmental Research Genomics: GTL program (grant number DE-FG02–04ER63931), and to Mike Cary and Jeremy Zucker for careful review of the manuscript and valuable assistance.

References

1 Cary, M.P. *et al.* (2005) Pathway information for systems biology. *FEBS Lett.* 579, 1815–1820
2 Krieger, C.J. *et al.* (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 32, D438–D442
3 Keseler, I.M. *et al.* (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* 33, D334–D337
4 Bader, G.D. *et al.* (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31, 248–250
5 Hermjakob, H. *et al.* (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.* 22, 177–183
6 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531
7 Paley, S *et al.* (1997) A Generic Knowledge-Base Browser and Editor. *AAAI97*.
8 Lemer, C. *et al.* (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.* 32, D443–D448
9 Overbeek, R. *et al.* (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28, 123–125
10 Kanehisa, M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280
11 Joshi-Tope, G. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 33, D428–D432
12 The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29
13 Weininger, D. (1988) SMILES, a chemical language and information system. *J. Chem. Inf. Comput. Sci.* 28, 31–36
14 Murray-Rust, P. and Rzepa, H.S. (2003) Chemical markup, XML, and the World Wide Web. 4. CML schema. *J. Chem. Inf. Comput. Sci.* 43, 757–772
15 Rumble, J., Jr *et al.* (2001) Reliable solubility data in the age of computerized chemistry. Why, how, and when? *Pure Appl. Chem.* 73, 825–829
16 Noy, N., and McGuinnes, D. (2001) *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford University (http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html)
17 Horridge, M. *et al.* (2004) *A Practical Guide to Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools*. Edition 1.0. The University Of Manchester.