BIOINFORMATICS ORIGINAL PAPERS Vol. 21 no. 24 2005, pages 4401–4407
doi:10.1093/bioinformatics/bti718

*Databases and ontologies*

# Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX

Lena Strömbäck* and Patrick Lambrix
Department of Computer and Information Science, Linköpings universitet, S-581 83 Linköping, Sweden

Received on December 16, 2004; revised on August 29, 2005; accepted on October 13, 2005
Advance Access publication October 18, 2005

**ABSTRACT**

**Motivation:** Analysis and simulation of pathway data is of high importance in bioinformatics. Standards for representation of information about pathways are necessary for integration and analysis of data from various sources. Recently, a number of representation formats for pathway data, SBML, PSI MI and BioPAX, have been proposed.

**Results:** In this paper we compare these formats and evaluate them with respect to their underlying models, information content and possibilities for easy creation of tools. The evaluation shows that the main structure of the formats is similar. However, SBML is tuned towards simulation models of molecular pathways while PSI MI is more suitable for representing details about particular interactions and experiments. BioPAX is the most general and expressive of the formats. These differences are apparent in allowed information and the structure for representation of interactions. We discuss the impact of these differences both with respect to information content in existing databases and computational properties for import and analysis of data.

**Contact:** lestr@ida.liu.se

## 1 INTRODUCTION

Currently, research within biology rapidly generates new knowledge on how genes, proteins and other substances interact. A complete description of the protein interaction network underlying cell physiology is seen as one of the major goals for proteomics by the Human Proteome Organization (Hermjakob *et al*., 2004). The US National Human Genome Research Institute (Collins *et al*., 2003) recognizes the understanding of genetic networks and protein pathways as crucial parts for two out of three thematic areas outlined for future genomics research. In particular, the understanding of how pathways contribute to the function of the cells and organisms, and the development of therapeutic approaches to diseases based on this knowledge are stated as two of the grand challenges for future research. They also recognize the development for reusable software modules, new ontologies (e.g. Lambrix, 2004) and improved technologies for database and knowledge management (e.g. Davidson *et al*., 1995) as means for finding solutions to these challenges in the future.

To fulfill this vision a format for representation of molecular pathways that allows for exchange, integration and easy creation of software tools is needed. Evaluations (Achard *et al*., 2001; McEntire *et al*., 2000) have shown that XML is an interesting and easy-to-use format for information representation and recently XML-based exchange formats for pathway information, e.g. SBML (Hucka *et al*., 2003), PSI MI (Hermjakob, 2004) and BioPAX (BioPAX working group, 2004, http://www.biopax.org), have been proposed.

The aim of this work is to compare and evaluate these proposals for new standards, to reveal their properties both as exchange languages and as general representation languages for pathway information. There are two main issues that we are interested in. The first is how well the formats can represent the main features of pathway data. The second issue is the ability of creation of software tools for data represented in this format. If the proposed standards meet the demands of these issues, it is very likely that they can provide a basis for future tools and reusable software modules.

The paper is divided into two parts. The first is a comparison of the type of information that can be represented within the three standards. The second part focuses on the two issues above. It contains a comparison with information in current databases and an investigation of possibilities for tools for export, import and analysis of data. The paper concludes with a discussion on important features for a standard for pathway representation.

## 2 STANDARDS FOR REPRESENTATION AND EXCHANGE OF PATHWAY DATA

In this section we present the three different standards, SBML, PSI MI and BioPAX. To allow a quick comparison of the standards Table 1 gives a summary of the main features for each of the formats. The structure of the table reflects the main properties of the standards that we compare and evaluate: the environment and usage of the standard; the representation of interactors, i.e. proteins and other molecules; the representation of the interaction; the possibility of representing other information in the standard; formal expressiveness and possibilities of referencing.

### 2.1 SBML

Systems Biology Markup Language (SBML) (Finney, 2004, http://www.cds.caltech.edu/~afinney/multi-component-species.pdf; Finney and Hucka, 2003, http://sbml.org/documents; Hucka *et al*., 2003) was created by the Systems Biology Workbench Development group in cooperation with representatives from many system and tool developers within bioinformatics. It is a language which aims to serve as a future standard for information exchange in computational biology and especially within molecular pathways. The aim of SBML is to model biochemical reaction networks, including cell signaling, metabolic pathways and gene regulation.

*To whom correspondence should be addressed.

Table 1. Main features of SBML, PSI MI and BioPAX

| | SBML | PSI MI | BioPAX |
|---|---|---|---|
| **Environment for the specification** | | | |
| Inventors | Systems biology workbench development group | Proteomics standards initiative | The BioPAX group |
| Existing tools | Tools for validation, visualization and conversion | Tools for viewing and analysis | The implementation in OWL can make use of existing tools such as Protégé |
| Used by | Used by around 75 systems, simulation and databases | Datasets available from IntAct, DIP and MINT. More databases, for instance BIND and HPRD, accept data in PSI MI | Collaboration with BioCYC, BIND and WIT |
| **Representation of interactors** | | | |
| Used notation | Species | Interactor | PhysEnity, with several subclasses |
| Description of parts of interactor | No current representation of parts of molecules but a proposal exists for next release of SBML | Protein sequences and sites can be described | Sequences adopted from PSI MI, molecule complexes and structure of small molecules |
| **Representation of interaction** | | | |
| Used notation | Reaction, modeling a transformation, transport or binding | Interaction | Interaction, with many subclasses, for instance, transport, catalysis and modulation |
| Role of interactor | Each reaction allows interactors of three predefined roles reactants products or modifiers | Each interactor can be given a role | Pathway representing a set of interactions |
| Number of interactors | Unbounded for each role | Unbounded number of interactors | Roles and number of interactors dependent on subtype |
| **Other predefined entities** | | | |
| Pathways | An SBML model encodes a reaction network | No representation of pathway or reaction networks | Specific data type for representation of pathways |
| Environment for interaction | Compartment defined as the environment for interactions | It is possible to define compartments for interactions | No environment for interactions |
| Experimental data | No data about experiments | Data about experiments verifying the interaction | No data about experiments |
| Mathematical relations | Mathematical relations for reactions | No mathematical relations | No mathematical relations, but details around interactions |
| **Expressiveness** | | | |
| Main structure | All entities defined on top level. References between them indicate the structure of interactions | Entities can be defined separately, but it is also possible to structure information around interactions | Entities are defined in an inheritance hierarchy |
| Inheritance | A hierarchy between the predefined entities but no possibility for the user to define types | No inheritance | Predefined inheritance hierarchy of subclasses for all entities |
| Definition of new attributes and entities | The note and annotation fields can be used for extra information | A specific list of attributes can be used to add information that does not fit into the format | The user is intended to use the concepts defined by the ontology, but it is possible to make application-specific additions |
| **Referencing to publications and databases** | References to other sources only in the annotation field | Links to publications and databases | Links to publications and databases |

The standard's main releases are called levels. Currently, level 2 is defined with a focus on models for analysis and simulation of basic biochemical networks. There is ongoing work on level 3 adding a number of features, for instance, model composition, description of molecule complexes, display and layout information and spatial characteristics of models.

SBML is widely used and reports that over 75 software systems use or convert to SBML. These systems include modeling and simulation of pathways; conversion to, for instance, Mathematica; drawing and visualization tools; and databases. Among the databases we can mention KEGG (Kanehisa and Goto, 2000), BioCyc (Karp *et al*., 2004) and Reactome (Joshi-Tope *et al*., 2005).

In the current SBML level each SBML model is intended to describe a biochemical network or a pathway. Each model contains a number of compartments, which is a description of the container or environment in which the reaction takes place. Compartments can be defined to be surrounded by each other. The substances or entities that take part in the reactions are represented as species.

```
<model name="Example">
<listOfCompartments>
  <compartment name="Mithocondrial Matrix" id="MM"/>
</listOfCompartments>
<listOfSpecies>
  <species name="Succinate"  compartment="MM" id="Succinate" />
  <species name="Fumarate" compartment="MM" id="Fumarate" />
  <species name="Succinate dehydrogenase"
            compartment="MM" id="Succdeh" />
</listOfSpecies>
<listOfReactions>
  <reaction name="Succinate dehydrogenas catalysis" id="R1">
    <listOfReactants>
      <speciesReference species="Succinate" />
    </listOfReactants>
    <listOfProducts>
      <speciesReference species="Fumarate" />
    </listOfProducts>
    <listOfModifiers>
      <modifierSpeciesReference species="Succdeh" />
      <modifierSpeciesReference species="S4" />
    </listOfModifiers>
  </reaction>
</listOfReactions>
</model>
```

**Fig. 1.** Example of SBML.

In SBML species can be everything from a simple ion, for instance a proton or an atom, through simple molecules, for instance glucose, to large molecules such as RNAs or proteins. For species it is possible to specify their spatial size and charge. It is also possible to specify model data, such as the initial concentration and amount and whether this can change during the reaction. The interactions between molecules are represented as reactions, defined as processes that change one or more of the species. The reaction can be a transformation, a transport or a binding reaction. Reactants, products and modifiers for reactions are specified by giving references to the relevant species. It is also possible to specify whether a reaction is reversible and to specify a reaction's speed by defining a kinetic law, mathematically describing the reaction. In addition to reactions, SBML also contains events, defined as discrete changes in the model. For an event it is possible to specify what triggers the event, time constraints and the result of the event. Finally, a model in SBML can also contain definitions of parameters, mathematical functions, units and mathematical expressions, which allows for shorter and more readable descriptions in the rest of the model. A simplified example of an SBML model, fetched from Reactome, is given in Figure 1.

In SBML the special fields note and annotation allow for addition of user- and software-specific information not contained in the rest of the standard.

## 2.2 PSI MI

The Proteomics Standards Initiative Molecular Interaction XML format (PSI MI) (Hermjakob *et al*., 2004a) was developed by the Proteomics Standards Initiative, one initiative of the Human Proteome Organisation (HUPO). The aim of the initiative is to develop standards for data representation in proteomics to facilitate data comparison, exchange and verification. One of those is the PSI MI standard for protein–protein interaction. The format is intended for exchange of data on protein interactions.

```
<entry>
<interactorList>
  <proteinInteractor id="Succinate>
     <names>
        <shortLabel>Succinate</shortLabel>
        <fullName>Succinate</fullName>
     </names>
  </proteinInteractor>
  ….
</interactorList>
<interactionList>
   <interaction>
     <names>
        <shortLabel> Succinate dehydrogenas catalysis </shortLabel>
        <fullName>Interaction between ....</fullName>
     </names>
     <participantList>
       <proteinParticipant>
          <proteinInteractorRef ref="Succinate"/> <role>neutral</role>
        </proteinParticipant>
       <proteinParticipant>
          <proteinInteractorRef ref="Fumarate"/><role>neutral</role>
       </proteinParticipant>
       <proteinParticipant>
          <proteinInteractorRef ref="Succdeh"/><role>neutral</role>
       </proteinParticipant>
     </participantList>
   </interaction>
</interactionList>
```

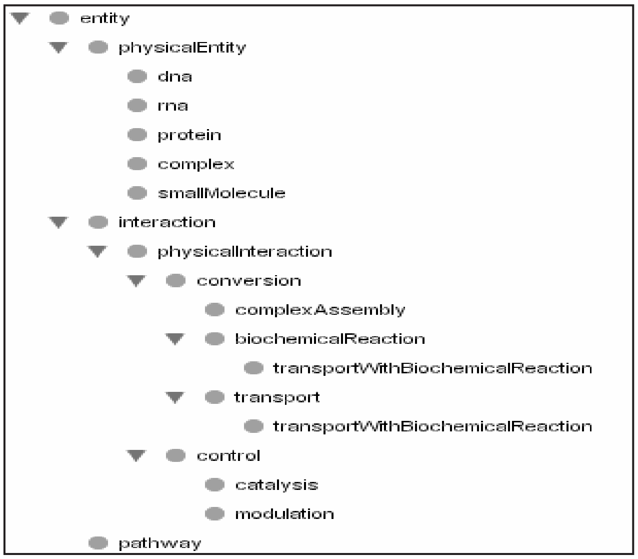
**Fig. 2.** Example of PSI MI.

PSI MI offers a number of tools for viewing and conversion of PSI MI data. In addition to this the tools Cytoscape and PIMWalker can be used for analysis of PSI MI data. Several databases, e.g. DIP (Salvinski *et al*., 2004), BIND (Bader *et al*., 2001), MINT (Zanzoni *et al*., 2002), IntAct (Hermjakob *et al*., 2004b) and HPRD (Peri *et al*., 2003), accept or export data in PSI MI format.

All data in PSI MI are structured around an entry. An entry describes one or more interactions that are grouped together for some reason. Note that a PSI MI model is not intended to be a pathway, an entry can be any set of interactions. In the entry the two tags source and the availabilitylist are used for describing the source of the data, usually an organization, and where the data can be accessed, typically a database. The experimentlist describes experiments and links to publications where the interactions are verified.

The pathway itself is described via the interactorlist, which is a list of proteins participating in the interaction, and the interactionlist, a list of the actual interactions. For each interactor information about, for instance, substructure can be defined. For each interaction it is possible to set the type of interaction and also a database reference to more information about the interaction. The type of the interaction, e.g. aggregation, is chosen from an externally defined controlled vocabulary, that can be chosen by the user. The participating proteins are described by their names or references to the interactorlist. It is also possible to set a confidence level for detecting this protein in the experiment, the role of the protein and whether the protein was tagged or overexpressed in the experiment. In addition each interaction has a description of availability and experiments which normally are references to the lists above.

Finally, the attributelist gives the user the possibility to add further information that does not fit into the entries above. Extra attributes can be used in all the parts described above. An abbreviated example pathway represented in PSI MI is shown in Figure 2.



- entity
  - physicalEntity
    - dna
    - rna
    - protein
    - complex
    - smallMolecule
  - interaction
    - physicalInteraction
      - conversion
        - complexAssembly
        - biochemicalReaction
          - transportWithBiochemicalReaction
        - transport
          - transportWithBiochemicalReaction
      - control
        - catalysis
        - modulation
  - pathway

**Fig. 3.** The BioPAX hierarchy.

## 2.3 BioPAX

The BioPAX Data Exchange format is defined by the BioPAX working group (BioPAX 2004). This group collaborates with several other efforts and databases. Examples of collaborators are Chemical Markup Language (Murray-Rust and Rzepa, 2002), SBML and CellML (Lloyd *et al*., 2004), BioCYC, BIND, Reactome and WIT. The aim of this standard is to define a unified framework for sharing pathway information. BioPAX is defined in a number of steps called levels. Currently, levels 1 and 2 exist. Level 1 focuses on metabolic networks and level 2 adds molecular interaction networks. Figure 3 shows the entity and interaction types allowed by these two levels. For the future there are plans for level 3, covering gene and DNA interactions, signal transduction and genetic interactions, and there is also a list of concepts for integration in later levels.

As the BioPAX definition and implementation are new, there are currently no specific tools for handling it or databases providing data for BioPAX. It is though possible to use Protégé (Noy, 2000) together with a specific plugin for viewing and editing BioPAX ontologies. There is also active work on conversion of data from several databases, such as BioCyc into BioPAX.

BioPAX makes more explicit use of relations between concepts than the other two standards and is defined as an ontology of concepts with attributes. It also provides an implementation in OWL. This makes it possible to benefit from reasoning and conclusions based on the semantics given by OWL and the ontology, but the cost is a higher computational complexity for reasoning and integration of data. This will be further discussed in Section 3.2.

In BioPAX all objects are described in a class hierarchy with Entity as the most general class. The BioPAX hierarchy as it appears when loaded into Protégé is shown in Figure 3. Entity has three subclasses PhysEntity, representing the interacting objects, Interaction, representing the interactions and Pathway, representing a set of interactions that together form a pathway model. PhysEntity has five subclasses, complex, protein, DNA, RNA and smallmolecule, describing different kinds of objects that may interact. For interaction there is a large number of subclasses. For each subclass there are given roles and numbers for the possible interactors. For some of the interaction subtypes it is possible to define additional information specific to this reaction.

The pathway entity used in BioPAX is of particular interest since it allows for combining interactions into pathway descriptions. This concept is a tool for building knowledge out of a large set of interactions. The order of the reaction is defined in the already defined interactions. Another interesting feature of BioPAX is the possibility of providing cross-references, which is a way to unify concepts and entities between data sources containing the same or similar information about a biological phenomenon.

# 3 DATABASE REPRESENTATION AND INTEGRATION

In this section we evaluate the three standard formats for exchange of pathway information with respect to the two issues stated in the Section 1. First we compare them with the information structure and content in existing databases and second we look at their properties for allowing easy creation of tools for analysis and import of data.

## 3.1 Comparison with existing databases

There are two basic principles on how molecular interaction data are organized within databases. Some databases, for instance, KEGG (Kanehisa and Goto, 2000, Kanehisa *et al*., 2004) SPAD (Tateishi *et al*., 1995) and BioCYC (Karp *et al*., 2004) are pathway databases, i.e. the information is connected to pathways, often presented for the user as a picture or map, presenting a number of interacting subjects and their relations. Other databases, for instance DIP (Salvinski *et al*., 2004), MINT (Zanzoni *et al*., 2002) and BIND (Bader *et al*., 2001), are interaction databases where the information is centered around the interacting subjects or the interaction.

For pathway databases the KEGG pathway database is a good representative. It provides information on molecular and gene interaction. Pathway information is provided by a set of reference maps, describing general information about known pathways. The maps can be specified for different species. The maps are clickable and provide links to protein and gene information from other databases. BioCYC and SPAD are similar to KEGG in the sense that data are presented for the user in clickable maps over the pathway.

Pathway databases can easily be represented in SBML where a pathway corresponds to an SBML model. The ordering of reactions is inferred via the reactant and product roles. It is currently possible to retrieve data from KEGG and BioCYC in SBML.

Regarding interaction databases, BIND stores all information as interaction pairs while DIP and MINT store information on the interacting molecules as well as the interactions themselves. The information stored in these databases is similar to the information that can be expressed in PSI MI and it is also possible to retrieve data from DIP and MINT in the PSI MI format.

The BioPAX standard allows for both representation of interactions, i.e. the concept interaction, and for putting them together as pathways, i.e. the concept pathway.

It is often hard to get a more detailed comparison of the structure of existing databases with the standards. This is both because the internal format of a database is not public, but also because they sometimes contain solutions that are specific to the technology solution of these databases. Instead we decided to compare with

**Table 2.** Main features of XIN, BIND XML and KGML

| | XIN | BIND XML | KGML |
|---|---|---|---|
| **Representation of interactors** | | | |
| Used notation | Node | Objects | Entry |
| Description of parts of interactors | No description of parts of molecules | Detailed description of reacting substructure | No representation of parts of molecules |
| **Representation of interaction** | | | |
| Used notation | Edge, can be further specified by subclasses | Interaction pair | Relation and reaction |
| Role of interactor | To and from, specifies direction | No roles | No roles for relation, substrate and product used for reaction |
| Number of interactors | Two interactors | Two interactors | Unbounded for each role |
| **Other predefined entities** | | | |
| Pathways | No pathway descriptions | Intended for interaction pairs | A KGML description describes a pathway |
| Environment for reaction | No environment for interaction | No environment for interaction | No environment for interaction |
| Experimental data | No data about experiments | Detailed data about experiments verifying the interaction | No data about experiments |
| Mathematical relations | No mathematical relations | No mathematical relations | No mathematical relations |
| **Expressiveness** | | | |
| Main structure | All entities are definied on top level, references between them | Information is structured around interaction. (It is also possible to generate the information structured around other types.) | All entities are defined on top level, references between them |
| Inheritance | Possible to add classes to nodes and edges but no inheritance | No inheritance | No inheritance |
| Definition of new attributes and entities | It is possible to define attributes to nodes and edges | No possibility to define entities | No possibility to define new entity types |
| **Referencing to publications and databases** | Links to databases, references to publications in defined attributes | Links to publications and other databases | Links to other databases but not to publications |

exported data in proprietary XML formats available from some databases. In particular we have looked at the formats provided by KEGG (KGML), DIP (XIN) and BIND. Table 2 gives a summary and comparison of the main features for these formats structured in the same way as the table for the standards above. From this table we can see that the main structure is very similar to the proposed standards. There are though some important differences and we conclude the section with a discussion of these. A more detailed discussion of the differences is given in Strömbäck (2004).

The most interesting difference is in the representation of interactions. Here all the three proprietary exchange formats and the three standards differ in their representation. For interaction databases the representation provided by PSI MI is a good model. In the table we can see that this representation is similar to but more general than what is used in BIND. XIN differs in the manner that it provides roles representing the direction of an interaction. For pathway databases, we can see the SBML interaction as a good model. Here the representation is instead the chemical process with reactant products and modifiers, very similar to what is provided by KGML. Note though that KGML does not include the representation of modifiers. For BioPAX it is possible to choose between using more general concepts similar to PSI MI or to be more specific in the interactor roles, similar to SBML. Note though that an SBML interaction with modifiers is represented as a control interaction with a conversion interaction as one of the participating interactors.

This difference in representation of reaction is related to the main purpose of the formats, i.e. SBML as a representation that is intended for simulations and PSI MI as a representation of experimental data. This difference can also be seen in the ability to represent more detailed information about the reaction. In SBML such information is given by defining mathematical formulas, while in PSI MI it is represented by giving details around the actual experiments. Here BioPAX is again different, having attributes containing information such as cofactors, enthalpy and entropy changes connected to the relevant interaction types.

Considering the representation of interactors, KGML and XIN have the possibility to further specify the class of the interactor, which is similar to what is provided in PSI MI. While all three standards allow, or plan to allow, specifying which parts of the molecules that interact, the only proprietary format that allows this is BIND.

## 3.2 Import and analysis of data

For the creation of parsers and for data analysis the most basic operation to provide is to extract relevant information from the provided XML file. We have tested this both by supplying queries directly on the format and by testing integration of data into relational databases. We will illustrate our findings with a discussion around the query: 'Find all entities that are involved in the same reaction as Succinate.' For the XML-based standards, SBML and

PSI MI, we use XQuery, the proposed standard query language for XML, for the illustration:

```
SBML:
for $i in document(‘‘sbml.xml’’)//reaction
   [listOfReactants/speciesReference/@species=‘‘Succinate’’ or
   listOfProducts/speciesReference/@species=‘‘Succinate’’]
return   {$i/listOfReactants/speciesReference/@species}
         {$i/listOfModifiers/speciesReference/@species}
         {$i/listOfProducts/speciesReference/@species}

PSI MI:
for $i in document(‘‘psi_mi.xml’’)//interaction [participantList/
   proteinParticipant/proteinInteractorRef/@ref=‘‘Succinate’’]
return   {$i/interactionType/names/fullName}
         {$i/participantList/proteinParticipant}
```

This clearly illustrates that there is an important difference between SBML and PSI MI because of the difference in representation of interactions. For SBML the interactors within an interaction are given roles, reactant, product and modifier. The format for interactions given by PSI MI is more flexible, and in this case allows for a simpler query. If we on the other hand needed to query based on the roles of the participants in a reaction this would have been easy in SBML. Which of the formats is preferable is dependent on the need for details about reactions and reactants in a specific application or the representation used by the importing database.

If we instead turn to BioPAX, we need a tool capable of reasoning with inheritance if we want to fully capture the semantical meaning of a BioPAX file. There are today several proposals for such query languages, for instance OWL-QL (Fikes *et al*., 2003), proposed as a probable future standard. For our tests, we chose to use a similar but simpler query language nRQL (Haarslev *et al*., 2004), which is currently available as a plugin to Protégé and therefore available for BioPAX. The above example could in nRQL be specified in two ways:

```
(retrieve (?r ?y) (and (?x |Succinate| |PHYSICAL-ENTITY|)
   (?r ?x |PARTICIPANTS|) (?r ?y |PARTICIPANTS|)))
(retrieve (?r ?y) (and (?x |Succinate| |PHYSICAL-ENTITY|)
   (?r ?x |LEFT|) (?r ?y |RIGHT|)))
```

In the first case the query is specified for all types of reactions where Succinate is involved while the second only finds reactions of type conversion with Succinate as one of the left reactants. This kind of use is also supported by the use of duplicate roles in BioPAX, i.e. all reactants are specified both as a participant and with a reaction-specific role.

This shows the benefit of using OWL for implementation of a standard. The price paid is a higher computational complexity. It is, however, an interesting topic for future research to investigate the real need of representational complexity and reasoning within molecular pathways. Such an investigation would likely suggest ways to develop specialized tools for BioPAX keeping the benefits and avoid the computational complexity of using full OWL.

## 4 IMPORTANT FEATURES OF A STANDARD FOR PATHWAY REPRESENTATION

The comparison and evaluation of the three proposed standards, SBML, PSI MI and BioPAX shows that all the formats provide a general framework for pathway representation but that there are a number of features that are of importance for a standard for pathway representation:

*Identification of proteins between data sources.* Since one important goal for creating a standard is information exchange and integration of data between different data sources it is very important that a particular molecule or protein can be identified between different sources. There are two ways of offering this in a standard. One is to use ontologies, i.e. standard names or identification numbers. All the standards allow or recommend this but none of them enforce it. Another way is to use identification via links to other databases, which PSI MI and BioPAX currently offer.

*Representation of protein structure.* To fully understand how pathways are built and how proteins and other molecules interact it is necessary to represent the interacting substructure of molecules. Currently, PSI MI and BioPAX allows representation of protein, DNA and RNA sequences while there is a proposal on representation of complexes for level 3 of SBML. For the future, more extensive solutions are needed.

*The granularity of reactions and roles of the reactants.* Search on connections between molecules within the network of reactions is very important to gain new information for pathways. This means that the representation of reactions is the key for providing efficient searches. A representation with few types of reactions and the same roles in all reaction types, like in SBML, is preferable if the user wants to search connections independent of the reaction types. If, instead, the user is interested in investigating a particular kind of interaction the more detailed representation and fine-grained roles of reactants is preferable. BioPAX provides both options in parallel.

*Information for reasoning.* One important purpose for representing pathway information is simulation and reasoning about the reactions and the functionality of the pathway. This requires information about the speed and conditions of each reaction described in the represented pathway. In SBML this kind of information can be represented as mathematical relations. BioPAX contains detailed information about reactions, but no mathematical formulas. PSI MI does not contain this kind of information. Instead it provides detailed information about experiments verifying reactions.

*Pathway presentation.* For pathway databases, the presentation of a pathway for the user is of importance. A common way to do this is as a map or a drawing. This map can be automatically generated from the pathway information, provided as within SBML or BioPAX. Another way, like in KEGG, is that the database also stores information about the actual map. Currently none of the presented standards contains this option. It is, though, on proposal for level 3 of SBML.

*User-defined entities and attributes.* PSI MI and SBML provide limited ways for the user for adding data that do not fit into the predefined attributes. In BioPAX it is possible to add new classes to the hierarchy. Since user-defined constructions are against the idea of standards one important goal would be to find constructions that enable sharing of the standard while still allowing for extendibility for specific applications.

## 5 CONCLUSION

In this paper we have evaluated three proposed standards, SBML, PSI MI and BioPAX for representation of pathway data. The evaluation consists of two parts. First we compared the formats with current databases and second we investigated their import and analysis capabilities. Our first investigation shows that all the standards provide a good representation, PSI MI is stronger on experimental data while SBML is better on simulation-related properties. BioPAX provides the richest and most general representation of the three. The second investigation shows that the richer hierarchy of BioPAX, which is a benefit with respect to representation of data, has a price with respect to computational complexity.

All this gives interesting implications for future work. One line of work would be a deeper investigation on how existing tools for working with XML such as query languages and database generators can be used for achieving integrated databases and discovery tools for data provided in the above formats. Here it would also be interesting to investigate limitations of the formats and the scalability of XML. Another line of future work would be to make a deeper comparison of biological features, investigating how they can best be modeled in each of the standards.

## ACKNOWLEDGEMENTS

We thank Helena Loborg for valuable insights about biological issues and Vaida Jakoniene for comments on a previous draft. P.L. is a member of the EU Network of Excellence REWERSE (Sixth Framework Programme project 506779) and participates in the working group 'Towards a Bioinformatics Semantic Web'.

*Conflict of Interest:* none declared.

## REFERENCES

Achard,F. *et al*. (2001) XML, bioinformatics and data integration. *Bioinformatics*, **17**, 115–125.

Bader,G.D. *et al*. (2001) BIND—the Biomolecular Network Database. *Nucleic Acids Res*., **29**, 242–245.

BioPAX working group (2004), BioPAX—biological pathways exchange language. Level 1, Version 1.0 Documentation.

Collins,F.S. *et al*. (2003) A vision for the future of genomics research: a blueprint for the genomic era. *Nature*, **422**, 835–847.

Davidson,S. *et al*. (1995) Challenges in integrating biological data sources. *J. Comput. Biol*., **2**, 557–572.

Fikes,R., Hayes,P. and Horrocks,I. (2003) OWL-QL–a language for deductive query answering on the semantic web. *Technical report*. Knowledge Systems Laboratory, Stanford University, Stanford, CA.

Finney,A. (2004) Systems biology markup language (SBML) Level 3: proposal: multi-component species features. Proposal manuscript. March 2004 (April 2004).

Finney,A. and Hucka,M. (2003) Systems biology markup language (SBML) Level 2: structures and facilities for model definitions. June 28, 2003 (April 2004).

Haarslev,V., Möller,R., Van Der Straeten,R. and Wessel,M. (2004) Extended query facilities for racer and an application to software-engineering problems. In *Proceedings of the 2004 International Workshop on Description Logics (DL-2004)* Whistler, BC, Canada, June 6–8, 2004, pp. 148–157.

Hermjakob,H. *et al*. (2004a) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol*., **22**, 177–183.

Hermjakob,H. *et al*. (2004b) IntAct—an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.

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.

Joshi-Tope,G. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33** (Database issue), D428–D32. PMID: 15608231.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **28**, 27–30.

Kanehisa,M. *et al*. (2004) The KEGG resources for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.

Karp,P.D. *et al*. (2004) The *E coli* EcoCyc Database: no longer just a metabolic pathway database. *ASM News*, **70**, 25–30.

Lambrix,P. (2004) Ontologies in bioinformatics and systems biology. In Dubitzky,W. and Azuaje,F. (eds), *Artificial Intelligence Methods and Tools for Systems Biology*. Springer, Berlin, pp. 129–146.

Lloyd,C.M. *et al*. (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol*., **85**, 433–450.

McEntire,R. *et al*. (2000) An evaluation of ontology exchange languages for bioinformatics. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 239–250.

Murray-Rust,P. and Rzepa,H.S. (2002) Towards the chemical semantic Web. In collier, H. (ed) (Infonortics). *Proceedings of the 2002 International Chemical Information Conference*, Tetbury UK, October 2002, pp. 127–139.

Noy,N.F., Fergerson,R.W. and Musen,M.A. (2000) The knowledge model of Protege-2000: combining interoperability and flexibility. In *Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2000)*, October 2–6, 2000, Juan-les-Pins, France, pp. 17–32.

Peri,S. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*., **13**, 2363–2371.

Salvinski,L. *et al*. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res*., **32** (Database Issue), D449–D451.

Strömbäck,L. (2004) XML representations of pathway data: a comparison. In *Proceedings of the ACM SIGIR'04 Workshop on Search and Discovery within Bioinformatics*. July 29, 2004, Sheffield, UK.

Tateishi,N., Shiotari,H., Kuhara,S., Takagi,T. and Kanehisa,M. (1995) An integrated database SPAD (signaling pathway database) for signal transduction and genetic information. In *Proceedings of the Genome Informatics Workshop (GIW95)* Yokohama, Japan, December 11–12, pp. 160–161.

Zanzoni,A. *et al*. (2002) MINT: a molecular interaction database. *FEBS Lett*., **513**, 135–140.