REVIEW

A review of standards for data exchange within systems biology

Lena Strömbäck, David Hall and Patrick Lambrix

Department of Computer and Information Science, Linköpings Universitet, Linköping, Sweden

The rapid increase in experimental data within systems biology has increased the need for exchange of data to allow analysis and comparison of larger datasets. This has resulted in a need for standardized formats for representation of such results and currently many formats for representation of data have been developed or are under development. In this paper, we give an overview of the current state of available standards and ontologies within systems biology. We focus on XML-based standards for exchange of data and give a thorough description of similarities and differences of currently available formats. For each of these, we discuss how the important concepts such as substances, interactions, and experimental data can be represented. In particular, we note that the purpose of a standard is often visible in the structures it provides for the representation of data. A clear purpose is also crucial for the success of a standard. Moreover, we note that the development of representation formats is parallel to the development of ontologies and the recent trend is that representation formats make more and more use of available ontologies.

Received: June 15, 2006
Revised: November 30, 2006
Accepted: December 4, 2006

Keywords:
Cellular pathways / Data exchange / Molecular interactions / Ontologies / Standardization

1 Introduction

Recently, systems biology researchers have started to rapidly generate new experimental results on genes, proteins, and the complex interactions between these and other substances. Because of this, large quantities of data are available, and there is an urgent need for means to allow quick comparison, analysis, and integration of data produced by different researchers and research teams. This need has been recognized by several main institutes. In particular, a complete description of the protein interaction network underlying cell physiology is seen as one of the major goals of proteomics by HUPO [1]. The US National Human Genome Research Institute [2] identifies the understanding of how pathways contribute to the function of the cells and organisms as one of the grand challenges for future research. These institutes also focus on the development of reusable software modules, new ontologies and improved technologies for database, and knowledge management as means for finding solutions to these challenges in the future.

An important prerequisite for providing the desired technology is the ability to supply molecular pathways in a format that allows for exchange, integration, and easy creation of software tools. Evaluations [3, 4] have shown that XML (extensible Markup Language, http://www.w3.org/XML) is an interesting and easy-to-use format for information repre-

Correspondence: Dr. Lena Strömbäck, Department of Computer and Information Science, Linköpings Universitet, Linköping 581 83, Sweden
E-mail: lestr@ida.liu.se
Fax: + 46-13-282666

Abbreviations: AGML, annotated gel markup language; **BSML,** bioinformation sequence markup language; **CML,** chemical markup language; **HUP-ML,** human proteome markup language; **MIAME,** minimum information about a microarray experiment; **MAGE-ML,** microarray and gene expression markup language; **MIRIAM,** minimum information requested in the annotation of biochemical models; **OBO,** open biomedical ontologies; **PSI MI,** proteomics standards initiative molecular interactions; **SBML,** systems biology markup language; **XML,** extensible markup language sentation in biological applications. This is also shown by the numerous proposals for the use of XML-based standards for representation of information within systems biology [5].We have previously compared ([6], updated in [7]) the three standards: SBML (Systems Biology Markup Language, http://sbml.org) [8], PSI MI (Proteomics Standards Initiative Molecular Interactions, http://psidev.sourceforge.net/mi/rel25/) [1] and BioPAX (Biological Pathways Exchange, http://www.biopax.org) [9].

The aim of this work is to widen the perspective by giving a more complete picture of available standards within the area. The scope of the paper is XML-based standards, which means that we have focused on standards defined in XML or the XML-based ontology languages RDF (Resource Description Framework, http://www.w3.org/RDF) and OWL (Web Ontology Language, http://www.w3.org/2004/OWL) [10]. We also put this into context by discussing the relation to ontologies, as well as newly developed standards aimed at defining the minimal content.

The paper starts with a discussion of different kinds of standards and their relation to ontologies. We then focus on XML-based standards and give an overview of available standards, showing which kind of information can be used for and also how established they are today in terms of available datasets and tools. We then go into details and discuss how various interesting concepts, such as substances, interactions and pathways are represented in the standards. The paper ends with a discussion in which we summarize our experience with these standards, and give our view of what the most important properties are for a standard.

2 Standardization, standards, and ontologies

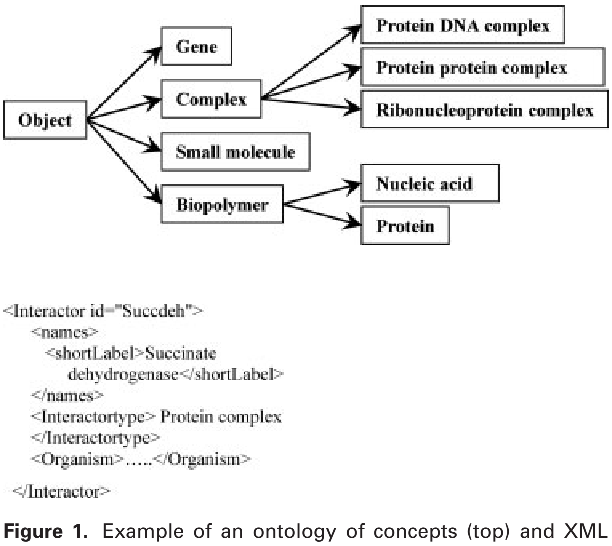
In this work, we focus on standards for efficient representation and exchange of information within systems biology, meaning that we are interested in standards for describing information about, for instance, substances, interactions between substances and pathway models where the aim is to find some common way of representing all these objects. There are basically two ways of addressing this problem: one is to start from the domain and define the different concepts, the other is to start from data originating from, for instance, experiments, and find models for representing this data.

When starting from the domain, the resulting formalization is often ontologies [11–14] of concepts. Intuitively, ontologies can be seen as defining the basic terms and relations of a domain of interest, as well as the rules for combining these terms and relations. Typically, the purpose of ontologies is to act as a community reference by providing a common terminology over a domain. The benefits of using ontologies include reuse, sharing, and portability of knowledge across platforms, and improved documentation, maintenance, and reliability. Ontologies lead to a better understanding of a field and to more effective and efficient handling of information in that field. A simple example of ontology is given in the upper part of Fig. 1. Here the arrows represent *IsA* relations, *e.g.* a *Protein IsA Biopolymer* meaning every protein is also a biopolymer. Ontologies differ regarding the kind of information they can represent but often the most important is the notion of concepts that represent the terms in a domain.

When starting from existing data on the other hand, the result is often different. A typical example of data represented in an XML standard for exchange is given in the lower part of Fig. 1. Here the protein complex *Interactor Succinate Dehydrogenase* is described. With regard to ontologies, concepts are important, but the most important task is to provide a structure in which different kinds of data can be efficiently represented and exchanged between researchers, *i.e.* to define a data model. As exemplified in the figure, representation formalisms such as XML are often used.

In reality, the distinction between ontologies and data representation formats is not always so clear. Ontologies can contain many kinds of relations and axioms to define further knowledge about concepts. Knowledge about the actual data items, instances, can be but is often not represented in ontologies. The most expressive formalisms in use for the representation of ontologies are description logic-based languages such as OWL. Regarding formalisms for data representation, many formats contain references to concepts defined in ontologies. In some cases, *e.g.* BioPAX, the concepts used within the format are defined as an ontology. Therefore, languages such as OWL can also be used in this case.

The development of ontologies and different kinds of representation formats is important for systems biology [2]. Biological ontologies have been around for a while and their use has grown drastically since data source builders concerned with developing systems for different (model) organisms joined to create the Gene Ontology Consortium in 1998 [15]. Currently, the field has matured enough to develop



```
<Interactor id="Succdeh">
    <names>
      <shortLabel>Succinate
         dehydrogenase</shortLabel>
    </names>
    <Interactortype> Protein complex
    </Interactortype>
    <Organism>.....</Organism>
 </Interactor>
```

Figure 1. Example of an ontology of concepts (top) and XML representation of actual data (bottom).

standardization efforts. An example of this is the organization of the first conference on Standards and Ontologies for Functional Genomics in 2002 and the development of the SOFG (Standards and Ontologies for Functional Genomics http://www.sofg.org) resource on ontologies. Also, OBO (Open Biomedical Ontologies http://obo.sourceforge.net) was started as an umbrella web address for ontologies that are in use within the biomedical domain. Many biological ontologies are already available *via* OBO. In parallel, formats for exchange of data have developed from being formats aimed at export of information from one particular tool or database towards standardized descriptions of how to represent information within a particular area. SBML, PSI MI, and BioPAX are good examples of this. Many of these standards make use of ontologies as a means of standardizing the concepts used within the standards. This can either be done by making references to existing ontologies or by specifying controlled vocabularies as an ontology that is part of the standard. Thus, the current development strives towards merging information developed from ontologies and formats for representation of information.

The latest development within systems biology also includes efforts to determine minimum requirements for a standard. MIAME (Minimum Information About a Microarray Experiment) [16] defines minimum requirements for microarray data and within the genomic technology society, several minimum requirements have been developed, such as MIAPE (Minimum Information About a Proteomics Experiment) [17] for proteomics data and MIRIAM (Minimum Information Requested In the Annotation of biochemical Models) [18] for models in systems biology. A molecular interaction module for MIAPE was in beta phase as of May 2006 (http://psidev.sourceforge.net/). One common theme among these requirements is a link to ontologies by the recommendation to store metadata according to controlled vocabularies instead of free-text. Other important requirements are inclusion of information about participating substances, which organisms they are collected from, and references to sources in the literature.

The main topic for this article is standards for representation of molecular interactions data and we include a discussion on ontologies where relevant. For our evaluation, we have taken the recommendations and requirements in the above minimum requirements specifications into consideration when evaluating the different standards. Since no minimum reporting requirement for protein interaction data was released at the time of writing, we have focused on common things like the way metadata is described when it comes to substances used, interactions referenced, and organisms for which the data are relevant.

3 An overview of available standards

As stated above, there has been an increasing interest in using XML or the XML-based languages OWL and RDF for the representation of information within this area and this is also our main interest in this article. Our first wide search for XML-based standards within system biology found 85 standards of varying levels of interest. From a biological point of view we have therefore concentrated on standards for describing molecular interactions or signaling pathways. In addition, we have also included standards for describing objects often included in standards for signaling pathways, that is, standards for describing proteins, DNA, genes or other substances, compartments, and experimental results. Regarding standards for experimental results, there is a large number of standards for describing experimental settings and details of a particular experiment. As these standards are numerous and often tailored to a particular experimental equipment, we have been forced to handle them only in a very general manner here.

It turns out that even with these restrictions the number of proposed standards is large. To further narrow the selection, we have used the following criteria: the standards must have been under recent development or use, they are referred to in more than one source, there are data available in the standard or there are tools available for manipulation of data in the standard. A first overview of our selection of standards is given in Table 1. For each of the standards, we show who is responsible for its development, the stated main purpose, and what kind of tools and data are available for the standard. It can, in particular, be noted that the selected formats range from formats in which the main purpose is to support data from one or a few tools or databases to formats that have been developed with the purpose of being a general standard.

To provide a quick overview of the content of each of the standards, we have summarized the main features of information content in Table 2. In this table, we note whether the standards give information on substances, interactions, pathways, compartments, organisms, or experiments. For each of these we denote, by using one or several letters, which kind of information the standard contains. Here, S means that the standard defines a preferred structure for the object, O that the standard provides an ontology of concepts for the objects, L that the standard provides means for linking to databases or other standard descriptions, and U that the object is used but with unspecified or no structure.

By studying this table, we note that the standards fall into three categories. First, we have the standards whose aim is to represent some aspects of molecular interactions or pathways. Here SBML and CellML [19] are tuned towards simulation, PSI MI towards experimental descriptions while BioPAX has a more general scope. The second group of standards aims at describing protein, DNA and RNA structure, and the third, experimental results.

Several of the formats contain ontology information or links to external sources. This use of external vocabularies is highly recommended in the specifications of minimal information for a standard. For instance, MIAME recommends [16] the use of controlled vocabularies and ontologies to represent data where such exist. MIAME uses the notion of

Table 1. Available standards, creators, and availability

Name	Ver.	Year	Defined by	Purpose	Tools	Data
SBML [8]	2.2	2003	Systems Biology Workbench development group	A computer-readable format for representing models of biochemical reaction networks	Supported by over 100 software systems	Data available from many databases, for instance, KEGG, www.genome.jp/kegg/ and Reactome, www.reactome.org
PSI MI [1]	2.5	2005	HUPO Proteomics Standards Initiative	A standard for data representation for protein-protein interaction to facilitate data comparison, exchange and verification	Tools for viewing and analysis.	Datasets available from many sources, for instance IntAct www.ebi.ac.uk/ intact/, and DIPdip.doe-mbi.ucla.edu/
BioPAX [9]	2	2005	The BioPAX group	A collaborative effort to create a data exchange format for biological pathway data	Existing tools for OWL such as Protégé protege. stanford.edu	Datasets available from Reactome www.reactome.org
CellML [19]	1.1	2002	University of Auckland and Physiome Sciences, Inc.	Support the definition of models of cellular and subcellular processes	Tools for publication, visualization, creation and simulation	CellML Model Repository (~240 models) www.cellml.org
CML [20]	2.2	2003	Peter Murray-Rust, Henry S. Rzepa.	Interchange of chemical information over the Internet and other networks	Molecular browsers, editors	BioCYC www.biocyc.org
EMBLxml [21]	1.0	2005	European Bio-informatics Institute	More stability and fine-grained modelling of nucleotide sequence information	API support in BioJavaX www.biojava.org	EMBL www.ebi.ac.uk/embl
INSD-seq [21]	1.4	2005	International Nucleotide Sequence Database Collaboration	The purpose of INSDSeq is to provide a near-uniform representation for sequence records	API support in BioJavaX www.biojava.org	EMBL www.ebi.ac.uk/ embl and GenBank www.ncbi.nlm.nih. gov/Genbank
Seq-entry	n/a	n/a	National Center for Bio-technology Information	NCBI uses ASN.1 for the storage and retrieval of data such as nucleotide and protein sequences. Data encoded in ASN.1 can be transferred to XML	SRI's BioWarehouse biowarehouse.ai.sri.com and Protein Structure Factory's ORFer www.proteinstruktur fabrik.de/orfer	Entrez www.ncbi.nlm.nih.gov/ Entrez
BSML [22]	3.1	2002	LabBook	Facilitate the interchange of data for more efficient communication within the life sciences community	LabBook's www.labbook.com Genomic Browser and Sequence Viewer. Converters	Previously provided by EMBL www.ebi.ac.uk/embl.
HUP-ML [23]	0.8	2003	Japan HUPO	A proteomics-oriented markup language for exchanging proteome data between researchers	HUP-ML Editor www1.biz.biglobe.ne.jp/ ~jhupo/HUP-ML/ hup-ml.htm	
MAGE-ML [24]	1.1	2003	Microarray Gene Expression Data	To facilitate the exchange of microarray information between different data systems	Converters	ArrayExpress www.ebi.ac.uk/ arrayexpress
MzXML [25]	2.1	2004	Institute for Systems Biology	The common file format for MS data	Converters, viewers	PeptideAtlas www. peptideatlas.org, Sashimi sashimi.sourceforge.net, Open Proteomics Database / apropos.icmb.utexas.edu/OPD
Mzdata [26]	1.05	2005	HUPO Proteomics Standards Initiative	To capture peak list information. Its aim is to unite the large number of current formats into one	Viewers, converters, analysis software, search engine	
AGML [27]	2.0	2004	Medical University of South Carolina	To model the concept of annotated gel (AG) for delivery and management of 2-DE results	Visualizer	AGML Central http:// bioinformatics. musc.edu/agml2/web/pages/ index.php

Table 2. Available standards, coverage and type of information

Name	Substances			Interactions	Pathways	Compart-ments	Organism	Experiments
	DNA, RNA	Protein	Other					
SBML	UL	UL	UL	SOL	SOL	SL		
PSI MI	SOL	SOL	SOL	SOL		L	SL	S
BioPAX	SOL	SOL	SOL	SOL	S		L	L
CellML	L	L	L	S	S	U	U	
CML			S	S				
EMBLxml	SL	SL					L	
INSDseq	SL	SL					L	
Seqentry	SL	SL					L	
BSML	SL	SL					L	S
HUP-ML	SL	SL					L	S
MAGE-ML	L	L					L	S
mzXML								SO
mzData								S
AGML							U	S

For each standard we state how objects can be specified by: S, data structure; O, an ontology; L, linking to other sources; and U, the object is used with unspecified or no structure.

qualifier, value, source-triplets to reference external knowledge. The source can be defined either by the user himself or reference a controlled vocabulary or external ontology. Also the MIRIAM requirement has an annotation scheme for external resources that requires the use of unique resources identifiers (URIs) to identify model constituents, such as model, compartments, reacting entity or reaction. These URIs are unique, permanent references to information about the particular object in that database or controlled vocabulary that are built up so they do not necessarily reflect the current server address or entry name but contain information to identify organization, database, and accession code. In the standards listed in the table, the link or ontology concept all include some variant of this feature.

4 Representation of information

To be able to compare more thoroughly the different standards, we now take a closer look at which kind of information can be defined for each standard. We have chosen to structure the description according to the important objects in the above table. Therefore, we discuss representation of substances, interactions and pathways, experiments, and finally organism and compartment information. For each of these, we discuss similarities and differences between the standards.

4.1 Substances

By substances, we mean the fundamental objects that participate in interactions within living organisms. A substance can, in principle, be everything from small molecules through macromolecules such as proteins, DNA, and RNA, to complexes of macromolecules. As different formalisms have different scopes, there are significant differences between the kinds of molecules they aim to cover in their description and in the granularity of the descriptions. For instance, SBML, whose main aim is to describe mathematical models for simulation, currently puts little effort into describing substance types and experimental evidence. To get a view of the minimum requirements for the description of substances, the MIAPE MSI module (http://psidev.sourceforge.net/gps/miape/MIAPE_MSI_0.4.comp MCP.pdf), for example, specifies that the following information be included (among other things): database queried, accession code for identified proteins, and peptide sequences.

The difference in scope between standards reflects the information content in formalisms. We exemplify this by describing which kind of information can be given for a substance in SBML, PSI MI, and CellML respectively. In SBML, a substance is described by the structure *Species* and represents a chemical entity that can take part in a reaction. In PSI MI, the corresponding structure *Interactor* describes a molecule that takes part in an interaction. CellML uses a metadata framework (http://www.cellml.org/specifications/metadata) [28] for description and references of components that can also be used for describing interacting substances. Table 3 shows which information can be given for substances in these three standards.

For PSI MI and CellML, there is an indentation denoting that this information is given as a substructure to the attributes *organism* and *variable*. We have also aligned information that corresponds to the structures so that it is easy to see where there is an overlap between the formalisms and where

Table 3. Species information in SBML, PSI MI, and CellML

	SBML: Species	PSI MI: Interactor		CellML:component	
1	id	id		name	
2	name	names		dc:title dcterms:alternative	
3		xref		cmeta:bio_entity	
4	speciesType	interactortype			
5		organism			
6			ncbiTaxId		
7			names	cmeta:species	
8			celltype		
9	compartment		compartment	(∼ group)	
10			tissue		
11				cmeta:sex	
12		sequence			
13				variable	
14	initialAmount				initial value
15	initialConcentration				
16	substanceUnits				units
17	spatialsizeUnits				
18	hasonlysubstanceUnits				
19	boundaryCondition				
20	charge				
21	constant				

there are differences. From the table, it is clear that PSI MI provides many features for describing details about each *interactor,* such as its type, information about the *organism* it occurs in, and in the case of proteins or DNA, its *sequence.* To make information and naming consistent, PSI MI often provides the possibility to link to ontologies provided by other entities, for instance in the case of *cell type* or *tissue.* SBML, on the other hand, provides many features for describing information that is important for simulation, such as initial amount and concentration of the species. Note that the *speciesType* provided in SBML is not an external reference but an internal type used within the model. For CellML, some attributes use notation with a prefix followed by a colon, *e.g. dc: dcterms:* and *cmeta:*.

For those cases, the definition of the attribute is imported from another XML specification. In particular, the CellML metadata specification (*cmeta:*) contains information on *species, sex,* and so on about the object. Species information does not refer to any external source but the specification recommends the use of NCBI's (National Center for Biotechnology Information) Taxonomy Browser as resource for scientific names. The biological entity information is given as a reference to a database, such as UniProt or GenBank.

A fundamental difference in the choice of formalisms and descriptions is apparent if we compare the information given by PSI MI and BioPAX. As stated above, the scope of PSI MI is the description of molecular interactions, while the scope of BioPAX is a general framework for pathways. In general, the information content within BioPAX and PSI MI is similar; however there is a fundamental difference regarding how they use structure and ontologies.

The BioPAX hierarchy is shown in the left part of Fig. 2. In BioPAX, *Physical Entity* is the most general type for describing substances and it can be one of the following: a *complex, DNA, protein, RNA,* or *small molecule.* From the description, it is clear which properties or describing substructures are relevant for each class. If we, for instance, want to represent a protein we can use all the properties from *Physical Entity,* but in addition, give values for *sequence* and *organism.*

If this is compared to the situation for PSI MI, we can see that many of the attributes coincide with the attributes for BioPAX. The attribute *InteractorType* is a reference to the ontology to the right in Fig. 2, which is a part of the OBO (http://obo.sourceforge.net) ontologies. Here, many of the concepts have a clear correspondence to the ontology defined for BioPAX, but for PSI MI there are no constraints on which properties or attributes can be used for a particular kind of substance as there is for BioPAX. This makes it possible for the user to define erroneous substances, but it also gives him more freedom to combine the available constructions.

In addition to these formats which aim to represent interactions or pathways, there exist a number of formats for interchange of data about substances that can take part in an interaction. Chemical substances that are not a protein or contain nucleic acids are formalized by CML (Chemical Markup Language) [20]. For DNA, RNA, and protein sequences, there exists a number of formats, most of them

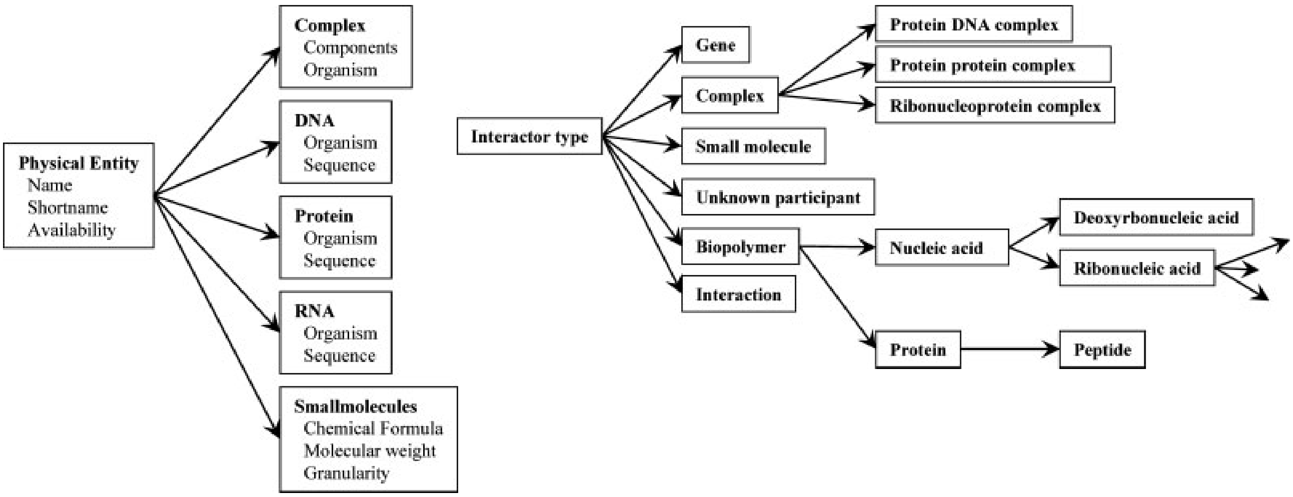


Figure 2. The BioPAX ontology (left) and the PSI MI (right) ontology for substances. Arrows represent *IsA* relations between concepts.

closely connected to existing databases (EMBLxml [21], INSDseq/INSDXML [21], Seqentry (http://www.ncbi.nlm.nih.gov/data_specs/dtd/NCBI_Seqset.mod.dtd)), or software tools (BSML (Bioinformation Sequence Markup Language) [22], HUP-ML (Human Proteome Markup Language) [23]). The sequence/feature formats typically contain one element with a listing of the nucleotide or amino acid sequence. The features are then given with a reference to the specific part of the sequence that it covers. Available formats are based on *de facto*-standard flat file formats such as FASTA and EMBL and are closely connected to the flat file formats they intend to replace. For higher level protein structures, there exist XML-based formats such as PDBML (Protein Data Bank Markup Language) [29], a format for representing data from The Protein Data Bank, and PSAML (Protein Structure Abstraction Markup Language) [30], which describes secondary structure elements and their relationships.

An interesting possibility for sequence information is the possibility to refer to sub sequences. PSI MI, BioPAX and many of the specific formats for protein, DNA, and RNA representation offer this possibility. This is accomplished by giving offsets (position) in the sequence. For PSI MI and BioPAX, references are made by stating numbers representing the start and end of the requested subsequence, for INSDseq a position in a sequence with a start and an end is given as character data, *e.g.* “1..307”.

4.2 Interactions and pathways

An interaction can be any kind of interference between two substances. Such interference can take many forms. For PSI MI, an interaction is something that is detected within an experiment, while for CellML and SBML, an interaction is a structure for defining mathematical relations needed for simulation. There is currently little information on minimal requirements for interactions. The MIRIAM [18] requirement also allows references for reactions but the molecular interactions module for MIAPE is in beta stage (http://psidev.sourceforge.net/).

We start with explaining the ontology defined for BioPAX and then relate the representations given by the other formalisms to these descriptions. The main structure and most important features for interactions within the BioPAX hierarchy are shown in the upper part of Fig. 3. Note in particular that BioPAX makes use of multiple inheritance.

BioPAX divides *interactions* into the main subtypes *control* representing an interaction in which one entity regulates or otherwise influences another, and *conversion* in which one entity is physically transformed into another. As with substances in BioPAX the possible features for each kind of interaction are given by type. In particular, we can note that for a general interaction the participating substances of the interaction are defined by the general term *Participant*, while for a conversion, we use the more specific participants *Left* and *Right,* indicating the order of the reaction. Similarly, the more specific features *Controller* and *Controlled* can be used for control interactions. Note also the attributes *Interactiontype* and *Controltype* that can be used to further specify the type of the interactions by referring to a controlled vocabulary.

This can be compared to how interactions are described in CellML, SBML, and PSI MI, which is summarized in Table 4. The PSI MI *Interaction* corresponds to *Interaction* in BioPAX. The PSI MI attribute *interactionType* refers to an external controlled vocabulary, which is a part of OBO (http://obo.sourceforge.net). This ontology is compared to the BioPAX hierarchy in the lower part of Fig. 3. PSI MI contains more features for referring to experimental results and data around actual experiments than the other standards. This is also clear from the features for each participant, where information about a participant’s biological role and its experimental role can be given. An interesting feature

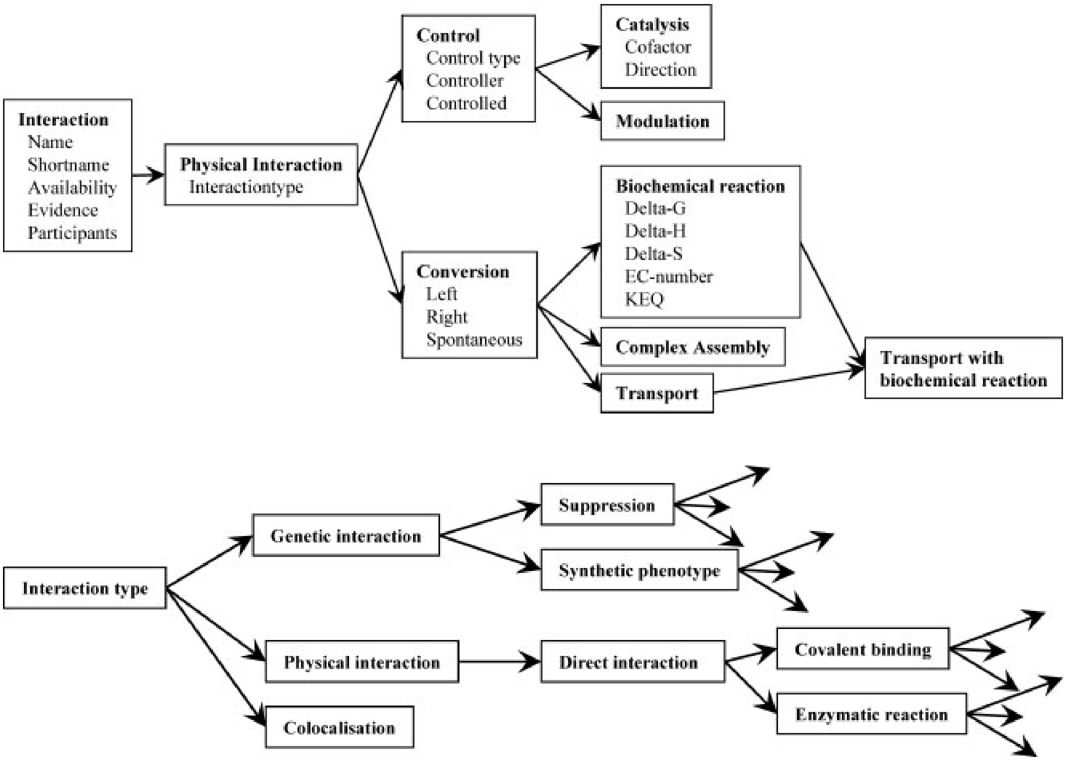


Figure 3. The BioPAX (top) and PSI MI (bottom) ontologies for describing interactions. Arrows represent *IsA* relationships between concepts.

is the inferred interaction list where the user can give information on which other interactions can be inferred from a given description.

For SBML, the concept reaction could be described as a combination of BioPAX' conversion and control, where *Reactant* and *Product* correspond to *Left* and *Right,* respectively, and *Modifier* is the *Controller,* and the *Controlled* item would be a reference to another reaction. Here *sboTerm* is a reference to an external controlled vocabulary SBO (Systems Biology Ontology), which is also part of OBO.

CellML also has the reaction concept but the description of a reaction is more complex than in SBML since both substance and reaction are represented by components that are connected by linking variables in the reaction component to variables in the participating substance components. So the *variable-ref* on line 9 in Table 4 is really a list of variables that in turn are connected to substances. The role for a substance in CellML (line 13) can be one of reactant, product, catalyst, activator, inhibitor, or modifier.

In addition to describing single interactions, it is often important to be able to combine interactions into larger networks, so-called pathways. In principle, a pathway can be any set of interactions that the user wants to group, but in most of the cases, we are interested in pathways where the interactions are connected through the participating substances. There is a clear difference between the standards and how they conceptualize this. Since one main aim of SBML is to describe mathematical models for pathways, a model in SBML is intended to describe a meaningful combination of connected interactions, thus a pathway. In CellML, interactions are connected to form a pathway by introducing a separate delta variable to the components representing substances for each reaction in which it participates. If a substance is the product of one reaction and reactant in another, the component representing the substance will have one delta variable to represent changes of concentration of the substance due to the first reaction and another delta variable for representing changes of concentration due to the second reaction. No other change in the two reactions is required. The main purpose of PSI MI, on the other hand, is often to record experimental results and interactions derived from these. From this perspective, a data set described in PSI MI does not normally describe a pathway; it can be any collection of interactions.

An extended means for describing pathways is given through the BioPAX pathway concept. With this construction, the user is able to group together any interactions he finds useful, and give the grouping a name that makes it easy for a user to reuse the information of particular interactions to create pathway models. The standard also allows representation of hierarchical pathways.

Table 4. Comparison of CellML component, SBML reaction, and PSI MI interaction

	CellML: component	SBML: Reaction	PSI MI: Interaction
1			imexID
2	name	id	id
3		name	
4			xref
5	variable		
6	reaction		
7		sboTerm	interactiontype
8			experimentList
9	variable-ref	reactant product modifier	participantList
10		id	id
11		name	names
12			experimental-role
13	role	sboTerm	biological-role
14			participantidentification
15			experimentalpreparation
16			confidencelist
17	direction		
18	delta_variable		
19	stoichiometry	stoichiometry	
20		kineticLaw	
21			inferredInteractionlist
22			participants
23			modeled
24			confidencelist
25	reversible	reversible	
26		fast	

4.3 Experiment descriptions and evidence

The ability to judge the validity of data is important within systems biology. In PSI MI, this is provided by a structure for describing experimental information. An experiment can be described by references to which kind of methods have been used, such as methods for detection of the interactions, the participants or specific features. Each of these can be specified by a controlled vocabulary in PSI MI. In addition, a measure of the confidence of the experiment can also be defined. On the other hand, BioPAX has a construction for describing evidence that can be connected to an interaction. It is similar to the PSI MI experiment construction, but in BioPAX the user can specify *confidence*, an *evidence code* or the *experimental* form of a participant in an experiment. Table 5 shows the main concepts for these constructions in PSI MI and BioPAX and how they relate to each other.

In addition to the experimental information in these two standards, specific data formats exist for a number of experimental methods. Some of these formats store not only the achieved results but also metadata on what experimental equipment and techniques were used during the experiment as well as subsequent analysis. We briefly mention a few of them here. For MS data, there are currently two formats, mzXML [25] and mzData [26]. mzXML was developed as a super format for existing proprietary formats for MS data from different manufacturers of MS equipment and mzData is an open standard format. There are plans to merge these two formats [31]. For microarray data, the format MAGE-ML (MicroArray and Gene Expression Markup Language) [24] can record all information about gene expression experiments required by MIAME. More general formats are HUP-ML [23] and AGML [27] (Annotated Gel Markup Language). HUP-ML was developed as an output format for J-HUPO's database. It is used to store data from proteome analysis with information about sample source, details of sample preparation, 2-DE images, spot identification, amino acid sequences, MS data, and so on. AGML can be used to store both gel and MS data as well as experimental methods.

Table 5. Experiment and evidence descriptions

	PSI MI: Experiment	BioPAX: Evidence
1	names	
2	hostOrganismList	
3	interactionDetectionMethod	
4	participantIdentificationMethod	
5	featureDetectionMethod	
6		evidence-code
7	confidenceList	confidence
8		experimental-form
9		participant

4.4 Compartments and species information

In most cases, it is important to record more information about where the interactions have been detected. Such information can have two purposes. When recording experimental results it is interesting to model which organism the molecules reside in and where the interactions occur, as this is a basis for how general the results are. For simulation purposes, it is important to model which substances occur together. In particular, for a transport interaction, the interaction is transporting molecules from one part of the organism, *e.g.* the cell, to another. In this case, it is important to describe different parts of the cells and how they are connected to each other. To this end, SBML contains the subclass compartment. A compartment is given an identity and properties describing its size. An important feature is the outside feature. With this the user can describe a structure between compartments by stating which compartment is outside the specified compartment. In CellML, components can be grouped together forming larger modules. Groups can specify physical or conceptual containment but lack the information on spatial dimensions that SBML offers.

For PSI MI and BioPAX, the focus lies more on recording of experimental results. Therefore, both these standards contain organism or species information, used for recording which organism this interaction has been detected for. Both formalisms also contain a compartment feature, in BioPAX called *Cellular-location*. In this case, however, the compartment is only specified by a name and a link to an external controlled vocabulary. Both standards recommend referring to Gene Ontology for the description of compartments.

5 Discussion

The overview of standards in section 3 of this article showed that there is a significant difference in scope between available XML-based formats for systems biology. In principle we could distinguish between three main aims: representation of molecular interactions and pathways, representation of substances, and representation of experimental data. Besides these three main aims there is also a difference in purpose of the formalisms, *i.e.* whether the standard is intended for the recording of results, models for simulation or something else. This purpose determines which terminology and sets of attributes are provided for every concept within a standard.

This is apparent for all the main elements in the standards. For representation of substances, there is a difference in granularity, *i.e.* how much information can be given for a concept, and which kind of substances can be represented for each of the standards. However, apart from this the representation is similar: *e.g.* two formalisms that are representing a protein structure do so in a similar way. The difference is greater when considering representation of interactions and pathways, as in this case the scope of the standard directly reflects which kinds of interactions and which kind of information it is relevant to represent. Also for representation of experimental results the difference is greater. This is due to the fact that the representation needs to range from representing some evidence for an interaction, through information about which kind of experiments have been performed, to detailed representation of results from a particular experimental method.

Even though this situation with many similar standards can be confusing for the user, the purpose of a standard is very important. Standards that have been created for a particular and well-defined purpose have often been more popular than general standards. This means that in the future, it is probable that there will be many standards with different scopes existing side by side and the user will often have to cope with handling data in more than one standard.

It is also very interesting to note the trend toward a merging of representation formats and the development of ontologies. There are two approaches: either the representation format is developed as an ontology, as is done for BioPAX or the format allows the user to use external ontologies. The first approach is beneficial if there is a need to represent many kinds of objects and different attributes for them, while the second approach is often more flexible and easy to use. This difference is also apparent if we consider representation languages for the approaches, where the technology for XML is currently more common and efficient [32, 33] than the technology for OWL.

Finally, we want to mention some implementation properties that can be of importance for the choice of representation format. One is the complexity of the XML tree structure, which can vary a lot between formats. For instance, BSML, INSDseq, and EMBLxml have less than 7 levels while Seqentry uses 26 levels for representing the same information. In most cases, this does not matter when working with tools built for the format but it makes manual processing and use of general purpose tools more difficult.

Another important feature is how well a format uses the XML structure. For instance, INSDseq has been developed as a common super format based on the flat file formats for GenBank, EMBL, and DDBJ. It has properties that deviate from the common XML way of storing data. For instance, position in a sequence is given as the character data “1..307” instead of as atomic data and in the *taxonomy* element all the data are stored in one element separated with semicolons. These are properties that make it hard to use available XML technology and the use and representation of these data as an XML structure would have been preferable.

To conclude, a user who is in need of a standard for representation should choose that standard based on how well the purpose of the standard coincides with his interest. The notes above are also important for a user who needs to design new tools for working with data. We would also recommend that a user makes use of standards that allow use of external vocabularies and links to other ontologies, as this kind of information is important for data integration, which is a prerequisite for increased understanding of the biological phenomena studied within the area.

We acknowledge the financial support of the Center for Industrial Information Technology. The third author is a member of the EU Network of Excellence REWERSE (Sixth Framework Programme project 506779).

6 References

[1] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J. *et al.*, *Nature Biotechnology*, 2004, *22*, 177–183.

[2] Collins, F. S., Green, E. D., Guttmacher, A. E., Guyer, M. S., *Nature*, 2003, *422*, 835–847.

[3] Achard, F., Vaysseix, G., Barillot, E., *Bioinformatics*, 2001, *17*, 115–125.

[4] McEntire, R., Karp, P., Abernethy, N., Benton, D. *et al.*, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2001, *8*, 239–250.

[5] Brazma, A., Krestyaninova, M., Sarkans, U., *Nat. Rev. Genet.*, 2006, *7*, 593—605.

[6] Strömbäck, L., Lambrix, P., *Bioinformatics*, 2005, *21*, 4401–4407.

[7] Strömbäck, L., Jakoniene, V., Tan, H., Lambrix, P., *Brief. Bioinform.*, 2006, *7*, 331–338.

[8] Hucka, M., Finney, A., Sauro, H. M. *et al.*, *Bioinformatics*, 2003, *19*, 524–531.

[9] Luciano, J., *Genome Technol.* 2006, 15–16.

[10] Wang, X., Gorlitsky, R., Almeida, J. S., *Nat. Biotechnol.*, 2005, *23*, 1099–1103.

[11] Lambrix, P., in: Dubitzky, W., Azuaje, F., (Eds.), *Artificial Intelligence Methods and Tools for Systems Biology*, Springer 2004, Chapter 8, pp. 129–146.

[12] Lambrix, P., Tan, H., Jakoniene, V., Strömbäck, L., in: Baker, C. J., Cheung, K.-H. (Eds.) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, Springer Verlag 2006, pp. 85–99.

[13] Staab, S., Studer, R., (Eds.) *Handbook on Ontologies*, Springer Verlag, 2003.

[14] Bodenreider, O., Stevens, R., *Brief. Bioinform.*, 2006, *7*, 256–274.

[15] Ashburner, M., Ball, C., Blake, J. A., Botstein, D. *et al.*, *Nat. Genet.*, 2000, *25*, 25–29.

[16] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G. *et al.*, *Nat. Genet.*, 2001, *29*, 365–371.

[17] Orchard, S., Hermjakob, H., Julian, R. K., Jr., Runte, K. *et al.*, *Proteomics* 2004, *4*, 490–491.

[18] Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S. *et al.*, *Nat. Biotechnol.* 2005, *23*, 1509–1515.

[19] Hedley, W. J., Nelson, M. R., Bullivant, D. P., Nielsen, P. F., *Philos. Transact. A Math. Phys. Eng. Sci.* 2001, *359*, 1073–1089.

[20] Murray-Rust, P., Rzepa, H. S., *J. Chem. Inf. Comput. Sci.*, 2003, *43*, 757–772.

[21] Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M. *et al.*, *Nucleic Acids Res.* 2006, *34*, D10–D15

[22] Cerami, E., *XML for Bioinformatics*, Springer Science, Business Media, New York 2005, pp. 17–47.

[23] Kamijo, K., Mizuguchi, H., Kenmochi, A., Sato, M. *et al.*, *Mass Spectrom. Soc.*, 2003, *51*, 542–549.

[24] Spellman, P. T., Miller, M., Stewart, J., Troup, C. *et al.*, *Genome Biol.*, 2002, *3*, research0046.1–0046.9.

[25] Pedrioli, P. G. A., Eng, J. K., Hubley, R., Vogelzang, M. *et al.*, *Nat. Biotechnol.*, 2004, *22*, 1459–1466.

[26] Orchard, S., Taylor, C. F., Hermjakob, H., Weimin-Zhu *et al.*, *Proteomics* 2004, *4*, 2363–2365.

[27] Stanislaus, R., Jiang, L. H., Swartz, M., Arthur, J., Almeida, J. S., *BMC Bioinformatics* 2004, *5*, 9.

[28] Cuellar, A. A., Lloyd, C. M., Nielsen, P. F., Bullivant, D. P. *et al.*, *Simulation* 2003, *79*, 740–747.

[29] Westbrook, J., Ito, N., Nakamura, H., Henrick, K. *et al.*, *Bioinformatics* 2005, *21*, 988–992.

[30] Kim, J.-H., Ahn, G.-T., Lee, M.-J., Lee, S.-H., *Proc. KORUS* 2003, *2*, 173–177.

[31] Orchard, S., Hermjakob, H., Taylor, C., Binz, P.-A. *et al.*, *Proteomics* 2006, *6*, 738–741.

[32] Strömbäck, L., Proc. 16th International Workshop on Database and Expert Systems Applications, IEEE Computer Society 2005, pp. 575–579.

[33] Strömbäck, L., Hall, D. in: by Grust, T., Höpfner, H., Illarramendi, A., Jablonski, S. *et al.* (Eds.) *Current Trends in Database Technology – EDBT 2006 Workshops*, LNSC 4254, Springer Verlag, 2006, pp. 220–233.