BioPAX for semantic web based data integration

Andrea Splendiani
Biomathematics and Biostatistics Dept.
Rothamsted Research
BioPAX for semantic web based data integration

Andrea Splendiani
Biomathematics and Biostatistics Dept.
Rothamsted Research
Outline

BioPAX overloading

BioPAX languages

OWL
RDF
XML

exchange language

data model

ontology

RDF-based data integration

roles of BioPAX in data integration

Network based data integration

Ondex
Network based data integration

- Cytoscape
- Ondex
Network based data integration

Ondex model:
- Network representing facts
- “Ontology” layer providing classes and properties
- Contexts
Network based data integration

Ondex
- Network representing facts
- "Ontology" layer providing classes and properties
- Contexts

Genes
Enzyme
Protein
React ion
Path ways
Envir onment

Network based data integration

- Data Input
- Data Integration
- Data Analysis

- Exchange formats: XGMML, RDF, OBO, PSI MI, SBML, FASTA
- Databases with OXL support: PHI-base
- Flatfile databases: KEGG, TF, TP, BioCyc, Drastic, MeSH, Medline

- Format importer
- OXL
- Flatfile parser

- ONDEX data integration framework
  - ONDEX core API
  - ONDEX Metadata
  - Ontology based graph structure
  - Exporter
  - Query API
  - Data alignment methods
  - Webservices
  - JSP webinterface
  - OXL

- FASTA
- RDF
- OXL
- Graph ML, GML
- XGMML
- SBML

- ONDEX Visualisation and Analysis Tool Kit (OVTK)
  - Taverna
  - Webinterface frontend

Network based data integration

What for?

Bioenergy crop improvement

http://www.biomass4energy.org
Network based data integration

- No a-priori boundaries
- Reconciliation

Ondex
Pathways from KEGG

More networks
Network based data integration

Extensional dimension: role of BioPAX

Search and visualize public biological pathway information. Single point of access. [more...]

Send us your feedback. Sign up for Pathway Commons announcements.

Search Pathway Commons:

To get started, enter a gene name, gene identifier or pathway name in the text box above. For example: p53, P38398 or mTOR.

To restrict your search to specific data sources or specific organisms, update your global filter settings.

Using Pathway Commons:

Biologists: Browse and search pathways across multiple valuable public pathway databases.

Computational biologists: Download an integrated set of pathways in BioPAX format for global analysis.

Software developers: Build software on top of Pathway Commons using our web service API. Download and install the cPath software to create a local mirror.

Current Data Sources:

Pathway Commons currently contains the following data sources:

- Cancer Cell Map, Release: 1.0 [19-May-06]
  - Browse

- HPRD [01-Sep-07]
  - Browse

- HumanCyc, Release: 10.5 [18-Sep-06]
  - Browse

- IntAct [14-Dec-07]
  - Browse

- MINT [21-Dec-07]
  - Browse

- NCI / Nature Pathway Interaction Database [28-Jan-08]
  - Browse

More networks
Network based data integration

Intensional dimension: role of BioPAX

BioPAX as an ontology for the integration of etherogenos networks
BioPAX overloading

- OWL
- RDF
- XML

Ontology

Data Model

Knowledge base: efficiency

PaxTools

Exchange language

Syntax, validation rules
BioPAX overloading

ontology

BioPAX is in part an ontology of biological networks

OWL

RDF

XML

Syntax, validation rules

Data Model

PaxTools
BioPAX overloading

ontology

- Entities, Interactions, their types and relations
- What do a “control” and a “genetic interaction share”? 
- What distinguishes them?

Ontology>>terminology!
BioPAX overloading

ontology

- Ontology definition:
  - An upper ontology as a reference.
  - Well defined high-level information (e.g. part of)
BioPAX overloading

ontology

ex. query: find all protein participating in a pathway
BioPAX overloading

ontology

ex. query: find all protein participating in a pathway
### BioPAX overloading

**part_of**

<table>
<thead>
<tr>
<th>ID</th>
<th>OBO_REL:part_of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>part_of</td>
</tr>
<tr>
<td>Relation properties</td>
<td>[transitive] [reflexive] [anti-symmetric]</td>
</tr>
</tbody>
</table>

**Definition**

For continuants: C part_of C' if and only if: given any c that instantiates C at a time t, there is some c' such that c' instantiates C' at time t, and c *part_of* c' at t. For processes: P part_of P' if and only if: given any p that instantiates P at a time t, there is some p' such that p' instantiates P' at time t, and p *part_of* p' at t. (Here *part_of* is the instance-level part-relation.)

**Comments**

Parthood as a relation between instances: The primitive instance-level relation p part_of p1 is illustrated in assertions such as: this instance of rhodopsin mediated phototransduction part_of this instance of visual perception. This relation satisfies at least the following standard axioms of mereology: reflexivity (for all p, p part_of p); anti-symmetry (for all p, p1, if p part_of p1 and p1 part_of p then p and p1 are identical); and transitivity (for all p, p1, p2, if p part_of p1 and p1 part_of p2, then p part_of p2). Analogous axioms hold also for parthood as a relation between spatial regions. For parthood as a relation between continuants, these axioms need to be modified to take account of the incorporation of a temporal argument. Thus for example the axiom of transitivity for continuants will assert that if c part_of c1 at t and c1 part_of c2 at t, then also c part_of c2 at t. Parthood as a relation between classes: To define part_of as a relation between classes we again need to distinguish the two cases of continuants and processes, even though the explicit reference to instants of time now falls away. For continuants, we have C part_of C1 if and only if any instance of C at any time is an instance-level part of some instance of C1 at that time, as for example in: cell nucleus part_of cell.

ex. query: find all protein participating in a pathway
BioPAX overloading

ontology

ex. query: find all protein participating in a pathway
BioPAX overloading

Data model

• Efficiency (space)
• Easy queriability (simple expressions)
• Represent information that is not ontologically clear (genetic interactions)
BioPAX overloading

Exchange language

- “Correct unit of information”/Validation.
- Necessity to be descriptive (an ontology is easily prescriptive!)

Knowledge base: efficiency

Data Model

OWL

XML

RDF

PaxTools

Syntax, validation rules

Exchange language

• RDF

• OWL

• PaxTools

• XML
## Languages in BioPAX

<table>
<thead>
<tr>
<th></th>
<th>Exchange</th>
<th>Data Model</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDF</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>OWL</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
# BioPAX overloading

Semantic Web metadata representation

<table>
<thead>
<tr>
<th></th>
<th>Exchange</th>
<th>Data Model</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDF</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>OWL</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
BioPAX and OWL

BioPAX, OWL, RDF

• BioPAX is represented in RDF
• BioPAX is serialized in XML
• The meaning of BioPAX is defined with terms defined in OWL (e.g.: disjoint class, necessary/sufficient restriction on properties,...)
• The use of OWL terms in BioPAX is restricted so that a pathway description is in OWL-DL.
• BioPAX doesn’t make use of the expressivity provided by OWL-DL
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

is biopax owl?
BioPAX and OWL

“participant” in BioPAX

<bp:biochemicalReaction rdf:ID="_2_H2O2____O2____2_H2O">  
  <bp:LEFT rdf:resource="#H2O2__peroxisomal_matrix_" />  
  <bp:RIGHT rdf:resource="#H2O__peroxisomal_matrix_" />  
  <bp:RIGHT rdf:resource="#Oxygen__peroxisomal_matrix_" />  
  <bp:EC-NUMBER rdf:datatype="http://www.w3.org/2001/XMLSchema#string">1.11.1.6</bp:EC-NUMBER>  
  <bp:NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2 H2O2 =>O2 + 2 H2O</bp:NAME>  
</bp:biochemicalReaction>
BioPAX and OWL

“participant” in OBO-OWL

<owl:Class rdf:about="http://purl.org/obo/owl/GO#GO_0000270">

...<owl:equivalentClass>
<owl:Class>
<owl:intersectionOf rdf:parseType="Collection">
<owl:Class rdf:about="http://purl.org/obo/owl/GO#GO_0008152"/>
<owl:Restriction>
<owl:onProperty>
<owl:ObjectProperty rdf:about="http://purl.org/obo/owl/OBO_REL#OBO_REL_has_participant"/>
</owl:onProperty>
<owl:onProperty>
<owl:someValuesFrom rdf:resource="http://purl.org/obo/owl/CHEBI#CHEBI_8005"/>
</owl:Restriction>
</owl:intersectionOf>
</owl:Class>
</owl:equivalentClass>
</owl:Class>
## BioPAX and OWL

### "participant" in RO

<table>
<thead>
<tr>
<th>ID</th>
<th>OBO_REL:has_participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>has_participant</td>
</tr>
<tr>
<td>Relation properties</td>
<td></td>
</tr>
<tr>
<td>inverse_of</td>
<td>OBO_REL:participates_in (instance level)</td>
</tr>
<tr>
<td>Definition</td>
<td>$P$ has_participant $C$ if and only if: given any process $p$ that instantiates $P$ there is some continuant $c$, and some time $t$, such that: $c$ instantiates $C$ at $t$ and $c$ participates in $p$ at $t$.</td>
</tr>
<tr>
<td>Comments</td>
<td>Has_participant is a primitive instance-level relation between a process, a continuant, and a time at which the continuant participates in some way in the process. The relation obtains, for example, when this particular process of oxygen exchange across this particular alveolar membrane has_participant this particular sample of hemoglobin at this particular time.</td>
</tr>
<tr>
<td>Sub-relation Of</td>
<td>OBO_REL:relationship</td>
</tr>
</tbody>
</table>
BioPAX and OWL

Properties and descriptions in BioPAX

OWL: properties are axioms, open world assumption
BioPAX and RDF

BioPAX in RDF

- BioPAX is essentially relying on RDF (RDFS)
- It is using OWL for “documentation” purposes
- It uses a subset of RDF that is compliant with OWL
• In principle it is possible to automatically classify entities or to discover that some class is equivalent to (or a subclass of) a biopax class

• In practice, types are usually known a-priori... given the number of classes involved, if they can be well described, classification may be unnecessary
BioPAX in RDF

Facts in RDF
<biopax-level2:protein rdf:about="#UniProt_P10599_Thioredoxin_ATL_derived_factor_ADF_Surface_associated_sulphydryl_protein_SASP_">
  <biopax-level2:COMMENT rdf:datatype="&xsd;string">
    FUNCTION: Participates ...
  </biopax-level2:COMMENT>
  <biopax-level2:SYNONYMS rdf:datatype="&xsd;string">TRDX</biopax-level2:SYNONYMS>
  <biopax-level2:SYNONYMS rdf:datatype="&xsd;string">TRX</biopax-level2:SYNONYMS>
  <biopax-level2:NAME rdf:datatype="&xsd;string">TRX</biopax-level2:NAME>
  <biopax-level2:SYNONYMS rdf:datatype="&xsd;string">TRX1</biopax-level2:SYNONYMS>
  <biopax-level2:SYNONYMS rdf:datatype="&xsd;string">TXN</biopax-level2:SYNONYMS>
  <biopax-level2:ORGANISM rdf:resource="#Homo_sapiens"/>
  <biopax-level2:DATA-SOURCE rdf:resource="#ReactomeDataSource"/>
  <biopax-level2:XREF rdf:resource="#UniProt_P10599"/>
</biopax-level2:protein>

<biopax-level2:unificationXref rdf:about="#UniProt_P10599">
  <biopax-level2:ID rdf:datatype="&xsd;string">P10599</biopax-level2:ID>
  <biopax-level2:DB rdf:datatype="&xsd;string">UniProt</biopax-level2:DB>
</biopax-level2:unificationXref>
BioPAX in RDF

Syntaxes: Turtle


BioPAX in RDF

Syntaxes: OWL/XML

```xml
<ClassAssertion>
  <OWLClass URI="&biopax-level2;protein"/>
  <Individual URI="&biopax;UniProt_P10599_Thioredoxin__ATL Derived_factor__ADF__Surface-associated_sulphydryl_protein___SASP_">
    ...
  </Individual>
</ClassAssertion>

<ObjectProperty URI="&biopax-level2;ORGANISM"/>
  <Individual URI="&biopax;UniProt_P10599_Thioredoxin__ATL Derived_factor__ADF__Surface-associated_sulphydryl_protein___SASP_">
    <Individual URI="&biopax;Homo_sapiens"/>
  </Individual>
</ObjectPropertyAssertion>

<ObjectProperty URI="&biopax-level2;XREF"/>
  <Individual URI="&biopax;UniProt_P10599_Thioredoxin__ATL Derived_factor__ADF__Surface-associated_sulphydryl_protein___SASP_">
    <Individual URI="&biopax;UniProt_P10599"/>
  </Individual>
</ObjectPropertyAssertion>

<DataPropertyAssertion>
  <DataProperty URI="&biopax-level2;COMMENT"/>
    <Individual URI="&biopax;UniProt_P10599_Thioredoxin__ATL Derived_factor__ADF__Surface-associated_sulphydryl_protein___SASP_">
      <Constant datatypeURI="&xsd;string">FUNCTION: ...
    </Individual>
</DataPropertyAssertion>

<DataPropertyAssertion>
  <DataProperty URI="&biopax-level2;NAME"/>
    <Individual URI="&biopax;UniProt_P10599_Thioredoxin__ATL Derived_factor__ADF__Surface-associated_sulphydryl_protein___SASP_">
      <Constant datatypeURI="&xsd;string">TRX</Constant>
    </Individual>
</DataPropertyAssertion>

<ClassAssertion>
  <OWLClass URI="&biopax-level2;unificationXref"/>
  <Individual URI="&biopax;UniProt_P10599"/>
</ClassAssertion>

<DataPropertyAssertion>
  <DataProperty URI="&biopax-level2;DB"/>
    <Individual URI="&biopax;UniProt_P10599"/>
      <Constant datatypeURI="&xsd;string">UniProt</Constant>
    </Individual>
</DataPropertyAssertion>

<DataPropertyAssertion>
  <DataProperty URI="&biopax-level2;ID"/>
    <Individual URI="&biopax;UniProt_P10599"/>
      <Constant datatypeURI="&xsd;string">P10599</Constant>
    </Individual>
</DataPropertyAssertion>
```
BioPAX in RDF

Facts in RDF
BioPAX in RDF

Meta-meta-view
CONSTRUCT ?x owl:sameAs ?y
WHERE {
  ?x bp:unificationXrefs ?x1 .
  ?x1 bp:ID ?i .
  ?x1 bp:version ?v .
  ?y1 bp:ID ?i .
  ?y1 bp:version ?v
}
CONSTRUCT ?interactor new:dirInt ?p
WHERE {
  ?x rdf:type bp:interaction .
  ?x bp:PARTICIPANTS ?px .
  ?px bp:PHYSICAL-ENTITY ?p
}
BioPAX in RDF

Interactions


[Define-interaction: (?pe1 new:interacts ?pe2) <- (?x rdf:type bp:interaction) (?x new:dirInt ?pe1) (?x new:dirInt ?pe2)]
BioPAX in RDF

Integration of BioPAX and other data sources
BioPAX in RDF

Relational declarative approach

### Chemical

<table>
<thead>
<tr>
<th>ID</th>
<th>Int</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASNumber</td>
<td>VARCHAR</td>
</tr>
<tr>
<td>Name</td>
<td>VARCHAR</td>
</tr>
<tr>
<td>ChebiID</td>
<td>Int</td>
</tr>
</tbody>
</table>

# Table Chemical

d2rq:dataStorage map:database;
d2rq:bNodesIdColumns "Chemical.ID";
d2rq:class biopax-level2:smallMolecule;

map:NAME a d2rq:PropertyBridge;
d2rq:belongsToClassMap map:chemical;
d2rq:property biopax-level2:NAME;
d2rq:pattern "@@Chemical.Name@@";
d2rq:datatype xsd:string;

map:CASNumber a d2rq:PropertyBridge;
d2rq:belongsToClassMap map:chemical;
d2rq:property phibase:CASNumber;
d2rq:pattern "@@Chemical.CASNumber@@";
d2rq:datatype xsd:string;

map:interaction_interaction_id a d2rq:PropertyBridge;
d2rq:belongsToClassMap map:chemical;
d2rq:property rdf:type;
d2rq:pattern "http://purl.org/obo/owl/CHEBI#CHEBI_@@Chemical.ChebiID@@";
### BioPAX in RDF

Direct mapping of tables to RDF

<table>
<thead>
<tr>
<th>CAS reg. Number</th>
<th>Chemical name</th>
<th>CHEBI ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>17804-35-2</td>
<td>Benomyl</td>
<td>3015</td>
</tr>
<tr>
<td>36734-19-7</td>
<td>Iprodione</td>
<td>28909</td>
</tr>
<tr>
<td>65277-42-1</td>
<td>Cis-Ketoconazole</td>
<td>47519</td>
</tr>
</tbody>
</table>

---

```
:ID1 rdf:type chebi:CHEBI_3015 ;
:CHEBI_3015 a biopax-level2:smallMolecule .
  biopax-level2:NAME "Benomyl"^^xsd:string .
  phibase:casNumber "17804-35-2" .

:ID2 rdf:type chebi:CHEBI_28909 ;
:CHEBI_3015 a biopax-level2:smallMolecule .
  biopax-level2:NAME "Iprodione"^^xsd:string .
  phibase:casNumber "36734-19-7" .

:ID3 rdf:type chebi:CHEBI_47519 ;
:CHEBI_3015 a biopax-level2:smallMolecule .
  biopax-level2:NAME "Cis-Ketoconazole"^^xsd:string .
  phibase:casNumber "65277-42-1" .
```
BioPAX in RDF

Ontologies as interpretation

Concepts in biopax can be used in different ontologies to provide interpretations of facts
BioPAX in RDF

BioPAX and Semantic Web based data integration

- **BioPAX**: concepts to describe pathways (entities, properties, abstraction (RDFS))
- **RDF**: (graph based) data model
- **RDF declarative processing of information** (sparql, rules)
- **Constraint checking** can be implemented as an additional layer
Re-composing BioPAX

Problems with pathway data integration

• Identity (beyond chemicals, states, generics, reactions)
• Representation dishomogeneity
• Non-explicitly encoded knowledge
Re-composing BioPAX

Problems with pathway data integration

```xml
<bp:catalysis rdf:ID="protein_tyrosine_kinase_activity_of_GP_VI_phosphorylated_Fc_Epsilon_R1_gamma_Collagen_IV__Syk_complex__plasma_membrane_">
  <bp:CONTROLLER rdf:resource="#GP_VI_phosphorylated_Fc_Epsilon_R1_gamma_Collagen_IV__Syk_complex__plasma_membrane_" />
  <bp:CONTROLLED rdf:resource="#Syk_mediated_phosphorylation_of_Phospholipase_C_gamma_2" />
  <bp:DIRECTION rdf:datatype="http://www.w3.org/2001/XMLSchema#string">PHYSIOL-LEFT-TO-RIGHT</bp:DIRECTION>
  <bp:CONTROL-TYPE rdf:datatype="http://www.w3.org/2001/XMLSchema#string">ACTIVATION</bp:CONTROL-TYPE>
</bp:catalysis>

<bp:biochemicalReaction rdf:ID="Syk_mediated_phosphorylation_of_Phospholipase_C_gamma_2">
  <bp:LEFT rdf:resource="#Phospholipase_C_gamma_2__plasma_membrane_" />
  <bp:RIGHT rdf:resource="#Phosphorylated_phospholipase_C_gamma_2__plasma_membrane_" />
</bp:biochemicalReaction>
```
Re-composing BioPAX

Problems with pathway data integration

```xml
<bp:catalysis rdf:ID="kinase_activity_of_PIP3_Phosphorylated_PKB_complex__plasma_membrane_"
    <bp:CONTROLLER rdf:resource="#PIP3_Phosphorylated_PKB_complex__plasma_membrane_" />
    <bp:CONTROLLED rdf:resource="#Phosphorylation_of_PDE3B" />
    <bp:DIRECTION rdf:datatype="http://www.w3.org/2001/XMLSchema#string">PHYSIOL-LEFT-TO-RIGHT</bp:DIRECTION>
    <bp:CONTROL-TYPE rdf:datatype="http://www.w3.org/2001/XMLSchema#string">ACTIVATION</bp:CONTROL-TYPE>
</bp:catalysis>

<bp:biochemicalReaction rdf:ID="Phosphorylation_of_PDE3B">
    <bp:LEFT rdf:resource="#PDE3B__cytosol_" />
    <bp:LEFT rdf:resource="#ATP__cytosol_" />
    <bp:RIGHT rdf:resource="#ADP__cytosol_" />
    <bp:RIGHT rdf:resource="#Phosphorylated_PDE3B__cytosol_" />
</bp:biochemicalReaction>

<bp:biochemicalReaction rdf:ID="Syk_mediated_phosphorylation_of_Phospholipase_C_gamma_2">
    <bp:LEFT rdf:resource="#Phospholipase_C_gamma_2__plasma_membrane_" />
    <bp:RIGHT rdf:resource="#Phosphorylated_Pholipase_C_gamma_2__plasma_membrane_" />
</bp:biochemicalReaction>
```
Conclusions

- Pathways as part of a broader set of biological information that can be represented as networks.
- BioPAX addresses distinct requirements, sometimes conflicting.
- Facts expressed in BioPAX can be seen as RDF. This is a good basis for network based data integration.
- Ontologies provide interpretations of these data.
Acknowledgments

BioPAX Working group

BioPAX-OBO

Ondex Team@Rothamsted:

- Chris Rawlings
- Jan Taubert
- Keywan Hassani-Pak
- Matthew Hindle
- Shaochih Kuo
- Artem Lysenko
- Mansoor Saqi
- Paul Verrier
- Catherine Canavet