
BioPAX

A Data Exchange Format for Biological Pathways

BioPAX Group
www.biopax.org

BioPathways SIG ISMB'03
Brisbane, Australia

BioPAX A Data Exchange Format for Biological Pathways

BioPAX Group

www.biopax.org

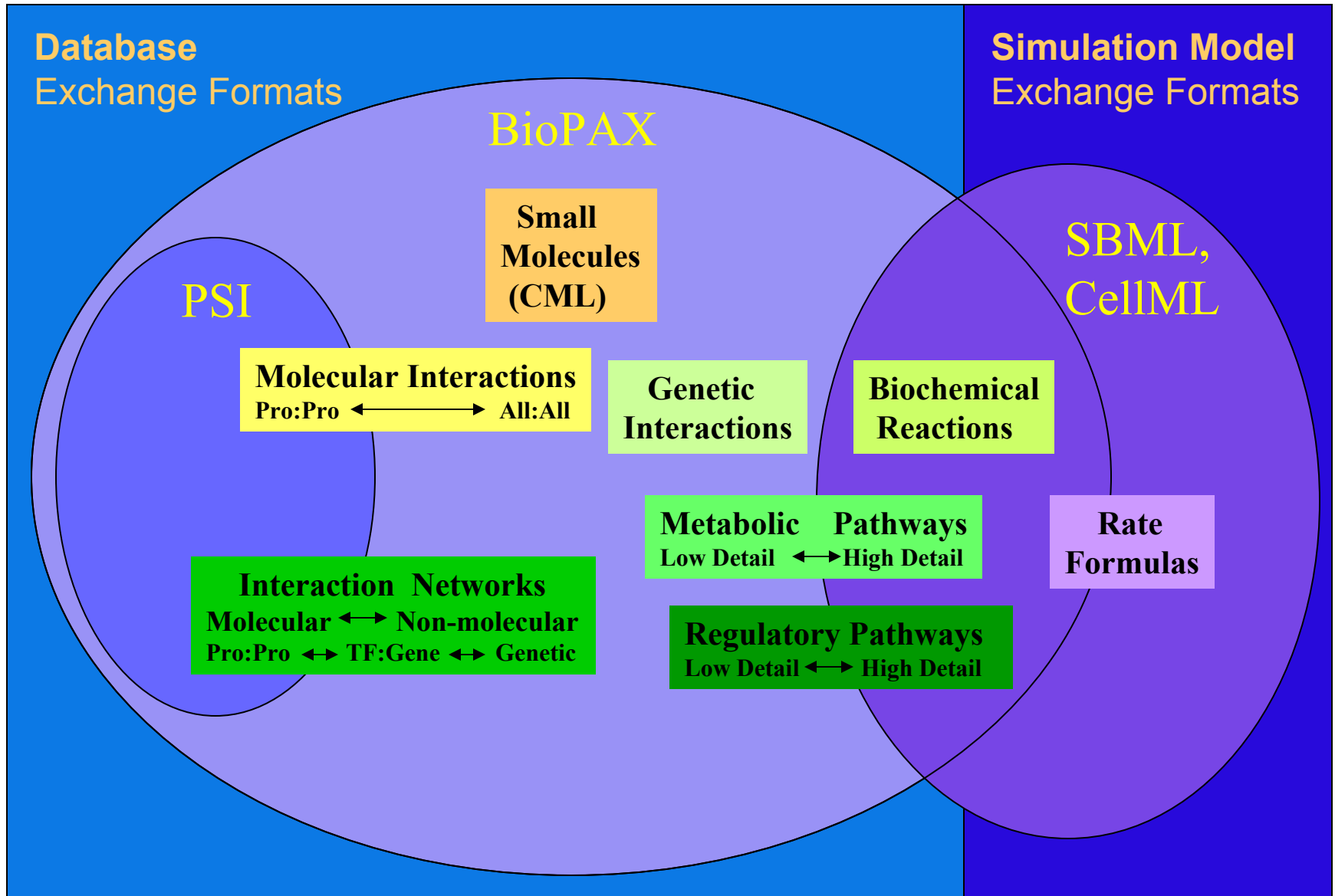
BioPathways SIG ISMB'03
Brisbane, Australia

Introduction



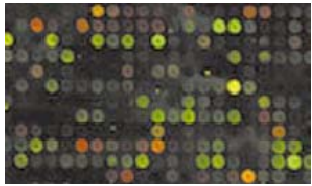
- BioPAX is a community-based effort conceived at ISMB '01; born at ISMB '02
- BioPAX = Biopathway Exchange Language
- A data exchange format intended to facilitate sharing of pathway data
- Provide a consistent format for pathway data to facilitate integration of pathway data from multiple sources.

Exchange Formats in the Pathway Data Space

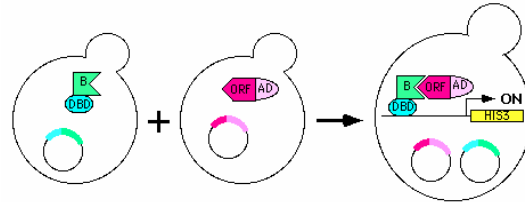


High Throughput Experimental Methods

Lots of data, lots of formats



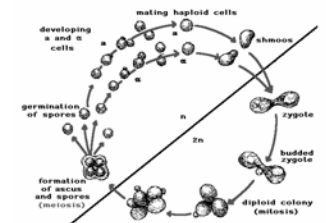
Microarray



Two-Hybrid



Mass Spectrometry

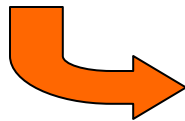


Genetics

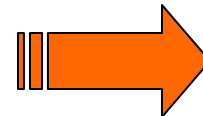
Expression, Interaction Data, Function, Protein modifications

Existing Literature

PubMed



Multiple Pathway Databases



Integration Nightmare!



Practical: Use Cases

- Build a centralized public pathway DB
- Share data between existing DBs
- Distribute proprietary data from a commercial enterprise
- Improve utility of pathway-based software
 - Expression analysis software
 - Pathway visualization tools
 - Simulation programs
 - Interaction, Network inference tools

Goals

- Accommodate existing database representations: BioCyc, BIND, WIT, aMAZE, KEGG, etc.
 - Compatible as a superset of representations
- Support different pathway types:
 - Metabolic pathways
 - Signaling pathways
 - Protein-protein interactions
 - Genetic regulatory pathways

Goals

- **Extensible:** Allow addition of new types of data
- **Encapsulation:** Represent an entire pathway in a single BioPAX record
- **Compatible:** Use existing standards
- **Flexible:** Allow different representations
- **Computable:** allow automated inference

Ontology Intro

- Natural language does a poor job at conveying complex information without ambiguity
- Ontologies provide a means to give concise meanings to pieces of data from a particular domain
 - Thereby facilitating computational operations on the data
- Ontologies are becoming increasingly common in the biological community
 - See <http://obo.sourceforge.net/obo.htm>

Ontology: Components*

- **Classes**
 - Arranged into a specialization hierarchy
- **Relations & attributes**
 - Fields (slots) on the classes that take values of specified types, types can be either basic (e.g. integer, string) or other classes
- **Constraints**
 - Define allowable values and connections within ontology
- **Objects & Values**
 - Together form instances of classes
 - Instances of BioPAX ontology will be created by users

* From Peter Karp, “Ontologies: Definitions, Components, Subtypes”, SRI International, presentation available at <http://www.biopax.org>

Ontology vs. Exchange Format

- Ontology and data exchange format (DEF) are not identical
- Ontology formalizes our pathway representation
- An ontology can be implemented as an exchange format
 - Multiple languages to choose from (XML, OWL, ASN.1, KIF, etc.)
 - Multiple ways to organize the data within each syntax

Choice of Language

- Currently translating BioPAX ontology into:
 - An XML Schema
 - Widely used syntax language
 - An OWL Ontology
 - More powerful data representation abilities
 - Community appears to be moving toward OWL (e.g. GO)
- Both are XML-based
- Both versions will be compatible with and fully translatable to each other
- Why two? Broad acceptance

BioPAX Ontology : Root

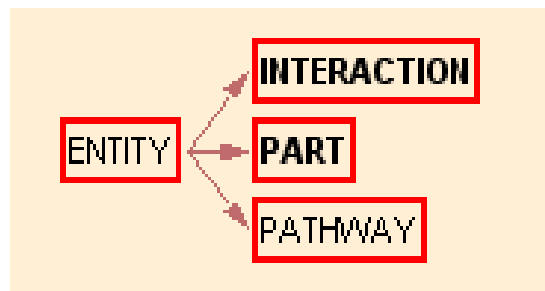
- Root class: Entity
 - Any concept that we will refer to as a discrete unit when describing the biology of pathways.
 - Does not include metadata
 - E.g. “DB source”, “PubMed ID”, “Experimental technique”, etc.



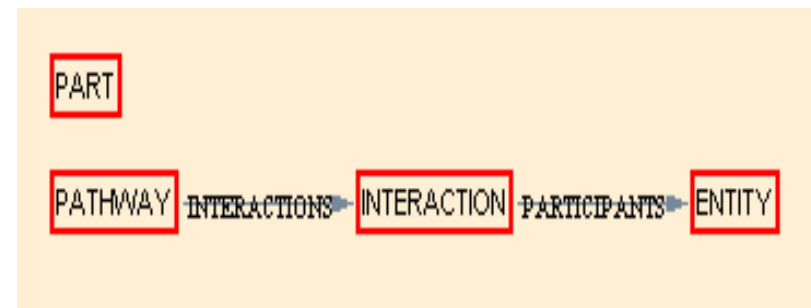
ENTITY

BioPAX Ontology : Root

- Entity Subclass: Part
 - A building block of simple interactions
 - E.g. Small molecules, Proteins, DNA, RNA
- Entity Subclass: Interaction
 - A set of entities and some relationship between them
 - E.g. Reactions, Molecular Associations, Catalyses
- Entity Subclass: Pathway
 - A set of interactions
 - E.g. Glycolysis, MAPK, Apoptosis



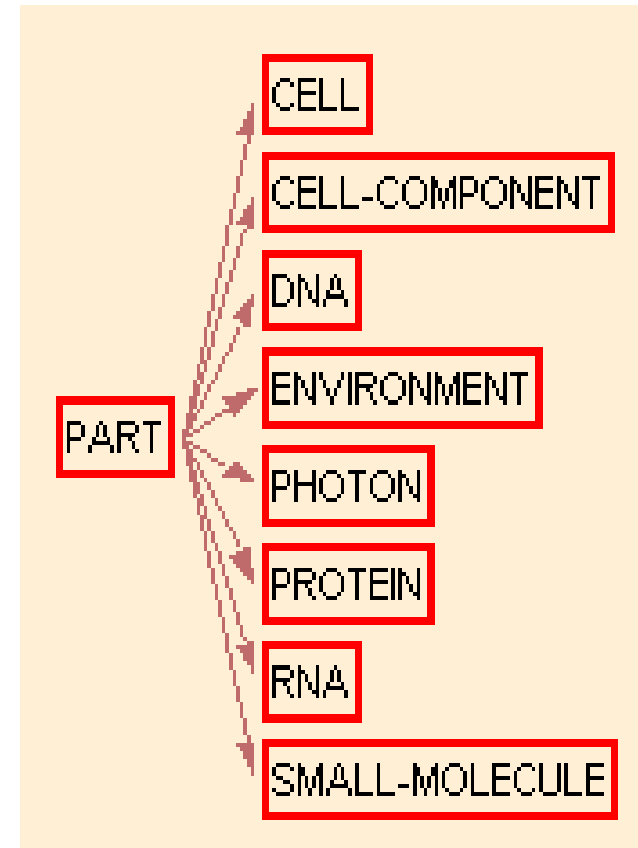
IS A



HAS A

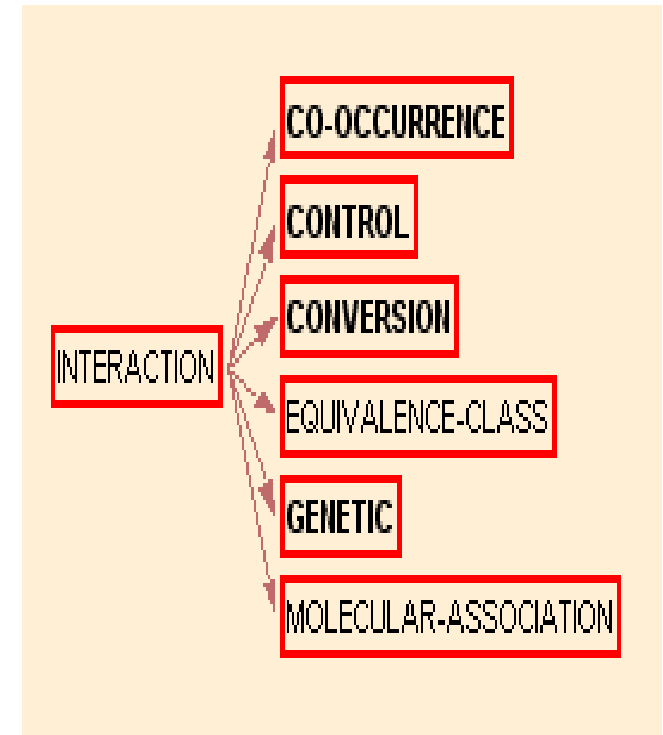
BioPAX Ontology: Parts

- **Cell**
 - A specific type of cell (e.g. cardiac myocyte, B lymphocyte).
- **Cell Component**
 - Part of a cell (e.g. nucleus, mitochondrion). The Gene Ontology contains a large list in the 'cellular component' ontology.
- **DNA**
 - Deoxyribonucleic acid (e.g. the EGFR DNA sequence; see GenBank for more examples).
- **Environment**
 - A physical or environmental effect (e.g. calcium wave, electric shock, heat, mechanical stress).
- **Photon**
 - Light at some intensity and wavelength (e.g. UV light).
- **Protein**
 - A protein (e.g. the EGFR protein sequence; see Swiss-Prot for more examples).
- **RNA**
 - Ribonucleic acid (e.g. messengerRNA, microRNA, ribosomalRNA)
- **Small Molecule**
 - A non-polymeric biomolecule. Generally, any bioactive molecule that is not a peptide, protein, DNA, RNA or possibly not a complex carbohydrate (e.g. glucose, penicillin)



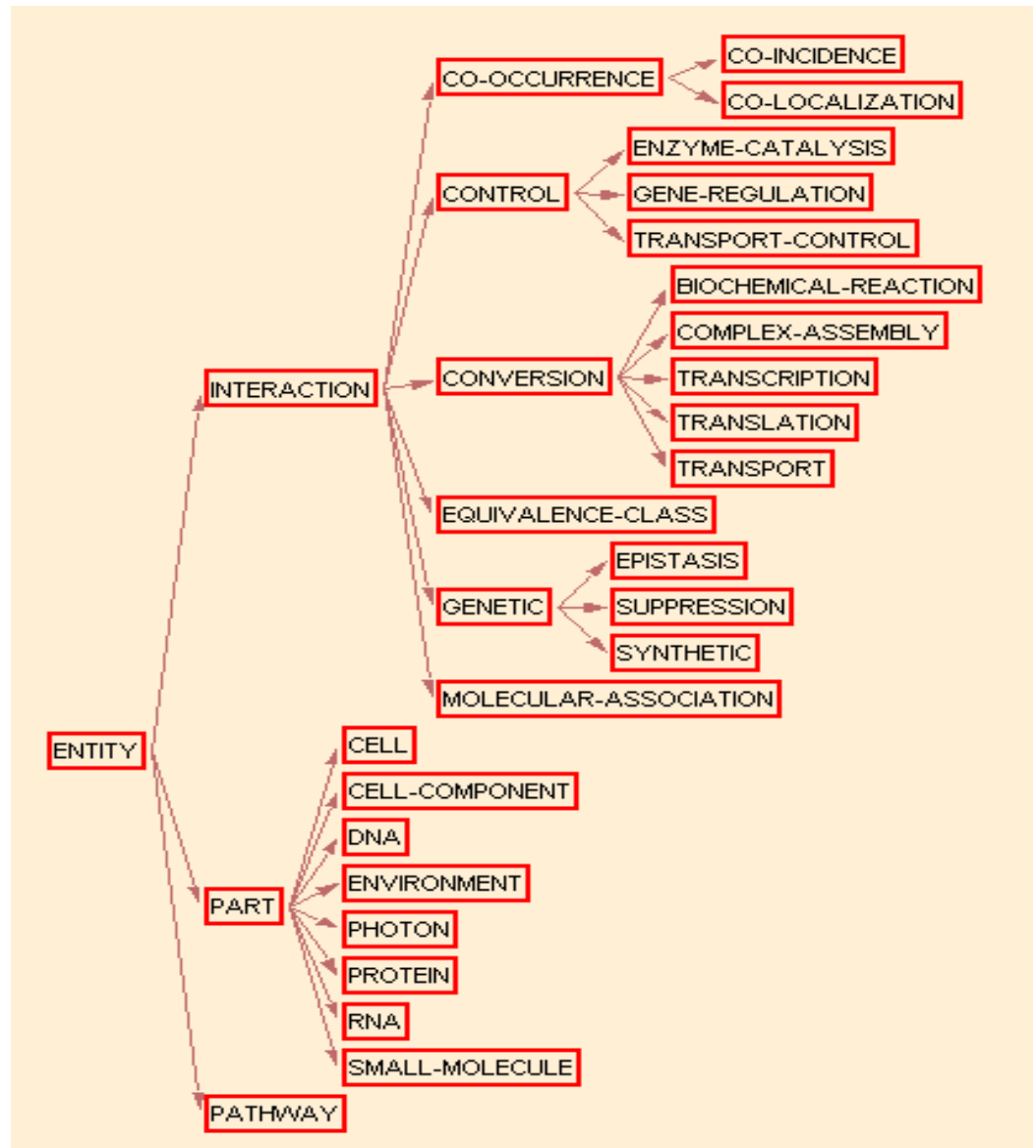
BioPAX Ontology: Interactions

- **Control**
 - The control of a process (e.g. enzyme catalysis controls a biochemical reaction, gene regulation controls gene expression).
- **Conversion**
 - A conversion process, which converts one set of entities to another set (e.g. a biochemical reaction converts substrates to products, the process of complex assembly converts single molecules to a complex, transport converts entities in one compartment to the same entities in another compartment).
- **Molecular Association**
 - An association between a set of molecules (e.g. Arp2-Arp3 protein-protein interaction; protein complex e.g. the result of a co-immunoprecipitation experiment; hexokinase-glucose).
- **Co-occurrence**
 - The co-occurrence of entities in some context. That context could be time, space, a sentence, sequence similarity space, etc. (e.g. Colocalization of a few receptors e.g. in a GPI anchored lipid raft; co-migration of cells; genes expressed at the same time).
- **Equivalence Class**
 - A set of entities that can be considered equivalent in some context (e.g. a set of paralogs that can replace each other as enzymes in a biochemical reaction, a set of enzymes that may not be homologs, but are functionally identical e.g. glucose-6-phosphatase).
- **Genetic**
 - A genetic interaction (e.g. a synthetic lethal interaction). An interaction between elements of a genotype that results in a change in phenotype.

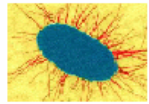


BioPAX Ontology

- Current class hierarchy
- Will be implemented in:
 - XML Schema
 - Widely used
 - OWL
 - Powerful data representation



Representing Metabolic Data in BioPAX

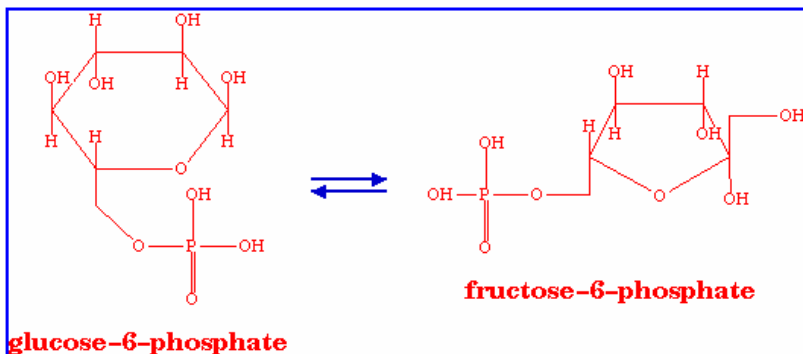


E. coli Reaction: 5.3.1.9

Superclasses: [5.3.1 -- INTERCONVERTING ALDOSES AND KETOSES](#)

[phosphoglucose isomerase](#): [pgi](#)

In pathway: [glycolysis](#), [gluconeogenesis](#)



ΔG° (kcal/mol): 0.4 [[1](#)]

Gene-Reaction Schematic: [?](#)



Unification Links: [ENZYME:5.3.1.9](#)

EcoCyc: Reaction



Reaction	
ID	1
Name	Glucose-6-p to fructose-6-p
Substrate	<cml>glucose-6-phosphate</cml>
Product	<cml>fructose-6-phosphate</cml>
Delta G	0.4 kcal/mole
EC	5.3.1.9

BioPAX: Reaction

Representing Metabolic Data in BioPAX (cont 1)

Enzymatic reaction of: phosphoglucose isomerase

Synonyms: glucose-6-phosphate isomerase , D-glucose-6-phosphate-ketol-isomerase

[glucose-6-phosphate](#) <=> [fructose-6-phosphate](#)

The reaction direction shown, that is, $A + B \rightleftharpoons C + D$ versus $C + D \rightleftharpoons A + B$, is in accordance with the direction of the reaction within a pathway.

In pathways: [gluconeogenesis](#) , [glycolysis](#)

Comment: 2-deoxyglucose-6-p is a known inhibitor in mammalian systems. E.coli cells with mutated *pgi* gene apparently utilize glucose primarily by the pentose phosphate pathway and to a lesser extent by the Entner-Duodoroff pathway. [3]

Citations: [3 , 4]

Inhibitors (neither competitive nor allosteric) [5]:

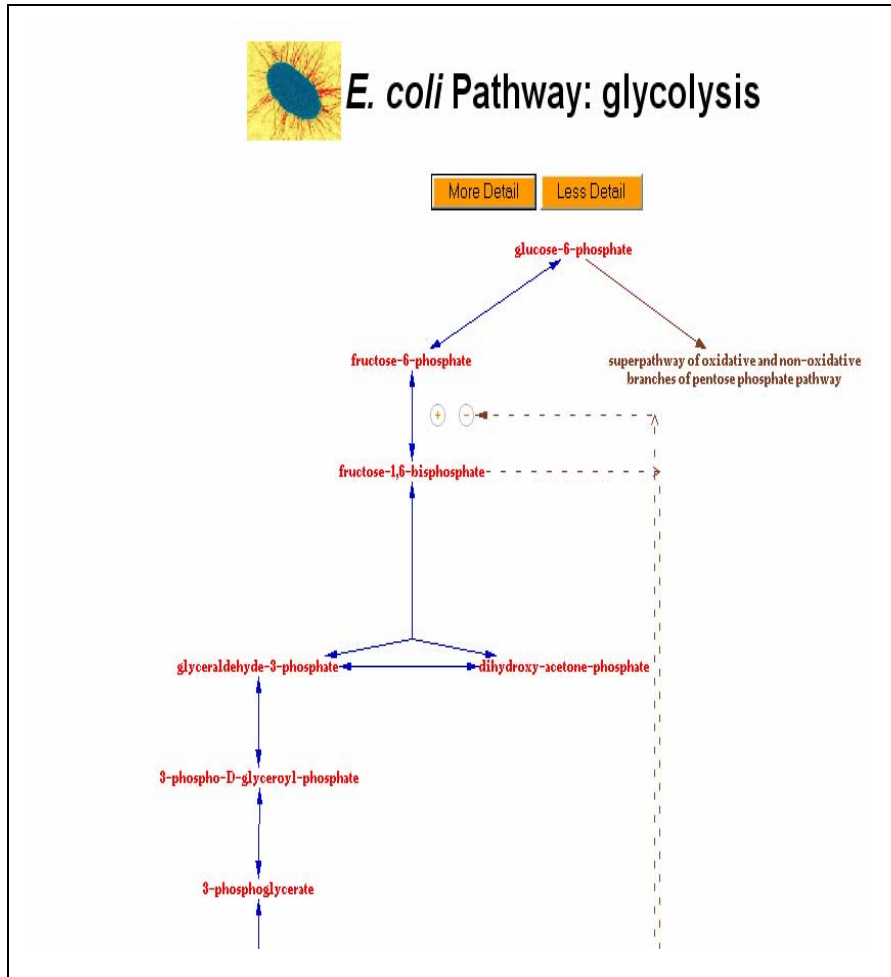


Catalysis	
ID	2
Name	Catalysis of glucose-6-p to fructose-6-p
Enzyme	glucose-6-phosphate isomerase
Reaction	BioPAX ID=1
Inhibitors	Low pH

EcoCyc: Enzyme-Catalyzed Reaction

BioPAX: Catalysis

Representing Metabolic Data in BioPAX (cont 2)



EcoCyc: Pathway



Pathway	
ID	10
Name	Glycolysis
Interactions	<ol style="list-style-type: none">1. BioPAX ID=22. BioPAX ID=43. BioPAX ID=6 etc.

BioPAX Class: Pathway

Converting PSI Data into BioPAX

```
<entrySet>
<entry>
<source>...</source>
<experimentList>...</experimentList>
<interactorList>...</interactorList>
<interactionList>
<interaction>
<experimentList>...</experimentList>
<participantList>
<proteinParticipant>
<interactorRef ref="hGHR"/>
<role>neutral</role>
<isTaggedProtein>>false</isTaggedProtein>
<isOverexpressedProtein>>true</isOverexpressedProtein>
</proteinParticipant>
<proteinParticipant>
<interactorRef ref="hGH"/>
<role>neutral</role>
<isTaggedProtein>>false</isTaggedProtein>
<isOverexpressedProtein>>true</isOverexpressedProtein>
</proteinParticipant>
</participantList>
<interactionType>
<names><shortLabel>binding</shortLabel></names>
<xref><primaryRef db="psicv" id="PSI:binding"/></xref>
</interactionType>
</interaction>
</interactionList>
</entry>
</entrySet>
```



Molecular Association	
ID	1
Name	hGHR binds to hGH
Participants	hGRH; hGH
DB Source	PDB:3HHR
Reference	PMID = 1549776
Experiment Description	X-ray Crystallography

PSI: Interaction (XML)

BioPAX: Molecular Association

Representing Signal Transduction in BioPAX

Cell_Signaling : MAP-kinase -> NF-kB

Entry Takako
From_molecule [MAP-kinase](#)
To_molecule [NF-kB](#)
Effect activation
Reference [\[Malinin_1997\]](#)
Role [apoptosis](#)

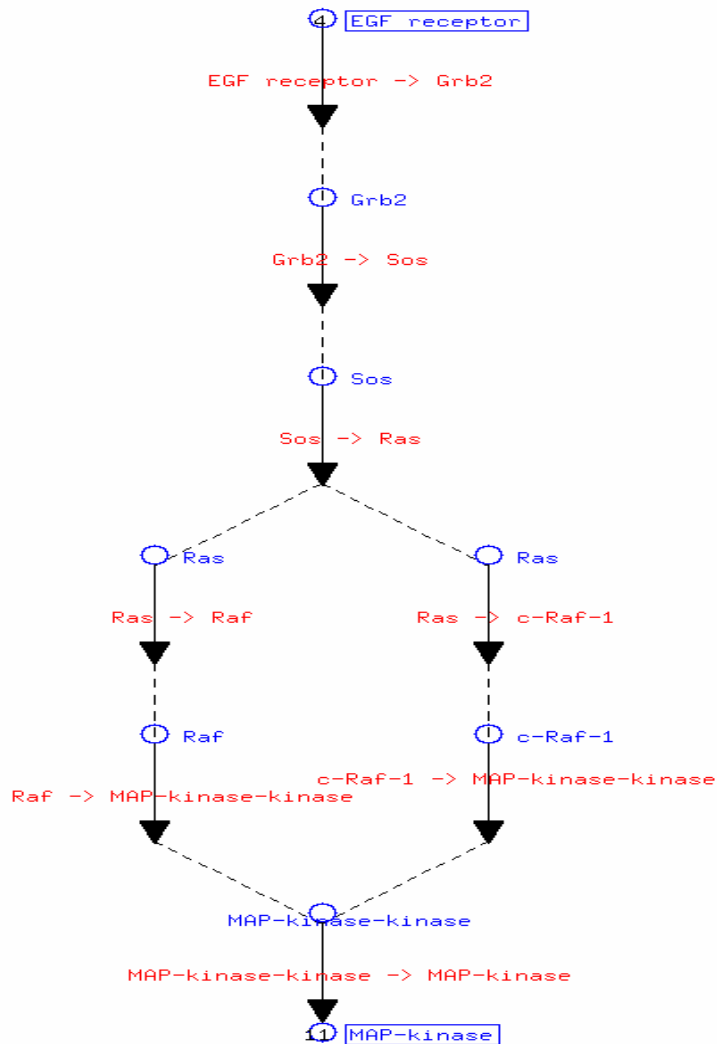
CSNDB Signaling Pathway Step



Reaction	
ID	20
Name	Activation of NF-kB
Substrate	NF-kB (inactive)
Product	NF-kB (active)

Catalysis	
ID	21
Name	MAP-kinase activates NF-kB
Enzyme	MAP-kinase
Reaction	BioPAX ID=20

Representing Signal Transduction in BioPAX



CSNDB Pathway



Pathway	
ID	10
Name	MAPK
Interactions	1. BioPAX ID=21 2. BioPAX ID=23 3. BioPAX ID=25 etc.

BioPAX Subgroups

- Created for multiple purposes:
 - Tackling specific conceptual problems
 - Developing spin-off projects
 - Small Molecule Database
 - Database of Pathway Resources
 - Gathering specific resources for core group
- Typically consist of:
 - Core group members (1-3)
 - Experts from external community (1-2)

BioPAX Subgroups: Small Molecule

- Evaluated CML 2.0 as means for exchanging small molecules
 - No comprehensive small molecule DB exists
 - Need to transfer entire small molecule structure, not just DB x-ref
 - Proof of concept:
 1. EcoCyc small molecules → CML 2.0 file
 2. CML 2.0 file → Shah lab visualization program
 - No loss of information

BioPAX Subgroups: States

- Determining best mechanism to represent biological states
 - E.g. post-translational modification states of proteins, cell-cycle states

BioPAX Subgroups: Examples

- Gathering sample data from various sources to illustrate use cases, promote practical development of BioPAX

Current Status

- Holding biweekly conference calls, bimonthly meetings
- Finishing Level 1 Ontology
 - Finishing slot definitions on Level 1 main-tree classes
 - Finishing class structure of side-trees
 - States, provenance, evidence, timing
- Working feverishly on presentation materials for ISMB 2003; presentation and poster will be available on the BioPAX web site.

Next Steps

- Finish level 1 ontology in GKB
- Implement ontology
 - In OWL (easy)
 - In XML Schema (slightly less easy)
- Export data from a few major DBs into BioPAX Level 1 XML files
 - Make revisions if necessary
- Release Level 1
 - By end of “summer” 2003 (hopefully)

GET INVOLVED!!!

- Participate in email list discussions to make your views heard
 - sign up via web site: <http://www.biopax.org>
- Join a subgroup
- Make your data available in BioPAX format, when complete
- Provide pointers to the BioPAX ontology from your database records to allow mapping between pathway databases
- POSTERS:
 - D-39 (database: BioPAX data exchange format)
 - K-39 (systems biology: BioPAX ontology)
- Birds of a Feather

BioPAX Supporting Groups

Groups

- **Memorial Sloan-Kettering Cancer Center:** Chris Sander, Gary Bader, Mike Cary, Joanne Luciano
- **University of Colorado Health Sciences Center:** Imran Shah
- **SRI Bioinformatics Research Group:** Peter Karp, Suzanne Paley, John Pick
- **BioPathways Consortium:** Eric Neumann, Vincent Schachter, Aviv Regev, Joanne Luciano (www.biopathways.org)
- **Argonne National Laboratory:** Natalia Maltsev
- **Samuel Lunenfeld Research Institute:** Chris Hogue
- **Harvard Center for Genomic Research:** Aviv Regev

Collaborating Organizations:

- **Proteomics Standards Initiative** (psidev.sf.net)
- **Chemical Markup Language** (www.xml-cml.org)
- **SBML** (www.sbml.org)
- **CeIIML** (www.cellml.org)

Databases

- **BioCyc** (www.biocyc.org)
- **BIND** (www.bind.ca)
- **WIT** (wit.mcs.anl.gov/WIT2)

Grants

- **Department of Energy**



Gene Network Reverse Engineering

Use Case

