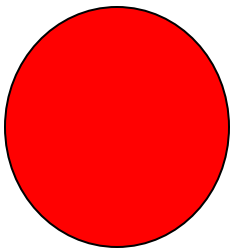


# BioPAX

## A Data Exchange Format for Biological Pathways

BioPAX Group  
[www.biopax.org](http://www.biopax.org)



# **Abstract**

**BioPAX (<http://www.biopax.org>) is a new community-based effort to develop a technical recommendation for a biological pathways data exchange format. This effort is timely as the number of new pathway databases being created is increasing and it is difficult to gather pathway information from these varied sources for analysis. A data exchange format will allow all participating databases to provide their data to users and to each other in a standard format, thus significantly reducing the amount of time spent on data integration by the bioinformatics community. The format is being designed to combine the strengths of existing biopathway databases such as BioCyc, BIND, WIT and aMAZE, among many others. In designing BioPAX, we endeavor to balance the many different representational needs of the biological pathways community by remaining flexible and extensible. A draft ontology will soon be available that provides a simple framework to be extended to include more detail via a leveled approach similar to that used by SBML. Encapsulation and compatibility are also being emphasized in the design and we are using existing standards when available. Currently, BioPAX has been designed to be compatible with PSI-MI (Proteomics Standards Initiative – Molecular Interactions - <http://psidev.sf.net>) and CML (Chemical Markup Language - <http://www.xml-cml.org>). An initial implementation of BioPAX will soon be available as both an XML Schema document and as an OWL ontology. The development of BioPAX is open for comment and feedback and participation is requested from all interested parties.**

# Introduction

- BioPAX – BioPathways Exchange
- Brief History
  - Why
  - Who
  - How
  - Where
  - When

# BioPAX Supporting Groups

## Groups

- Memorial Sloan-Kettering Cancer Center: C. Sander, J. Luciano, M. Cary, G. Bader
- University of Colorado Health Sciences Center: I. Shah
- SRI Bioinformatics Research Group: P. Karp, S. Paley, J. Pick
- BioPathways Consortium: J. Luciano, E. Neumann, A. Regev, V. Schachter ([www.biopathways.org](http://www.biopathways.org))
- Argonne National Laboratory: N. Maltsev
- Samuel Lunenfeld Research Institute: C. Hogue

## Collaborating Organizations:

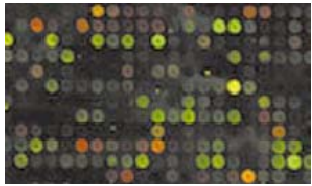
- Proteomics Standards Initiative ([psidev.sf.net](http://psidev.sf.net))
- Chemical Markup Language ([www.xml-cml.org](http://www.xml-cml.org))
- You?

## Databases

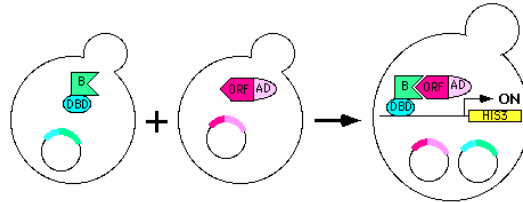
- BioCyc ([www.biocyc.org](http://www.biocyc.org))
- BIND ([www.bind.ca](http://www.bind.ca))
- WIT ([wit.mcs.anl.gov/WIT2](http://wit.mcs.anl.gov/WIT2))
- You?



# High Throughput Experimental Methods



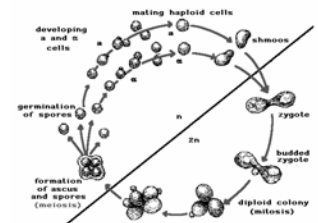
Microarray



Two-Hybrid



Mass Spectrometry



Genetics

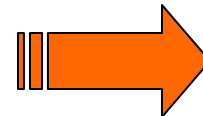
Expression, Interaction Data, Function, Protein modifications

Existing Literature

PubMed



Multiple Pathway Databases



Integration Nightmare!



# Motivations

- Facilitate exchange of pathway data
- Facilitate integration of existing pathway databases
- Allow pathway databases to exchange data in a common format
- Facilitate analysis of pathway data
- Allow pathway software to input and output data in a common format

# Goals

- Accommodate representations used in existing databases such as BioCyc, BIND, WIT, aMAZE, KEGG
- Include support for these pathway types:
  - Metabolic pathways
  - Signaling pathways
  - Protein-protein interactions
  - Genetic regulatory pathways

# Goals

- Programming language independent
- Utilize XML
  - XML Schema
    - Widely used
  - OWL
    - Powerful data representation

# Goals

- **Extensible:** Specific classes of data in BioPAX have been marked as extensible to allow addition of new types of data in the future
- **Encapsulation:** An entire pathway can be encapsulated in a single BioPAX record
- **Compatible:** BioPAX will try to use existing standards for encoding biological pathway related information wherever possible
- **Flexible:** Different preferred representations of pathway data can be described using BioPAX

# BioPAX Meetings

- Copenhagen 2001 (ISMB) – Open Pathways DB
- Edmonton, August 2002 (ISMB)
  - Pathways Repository focused effort on exchange standard
- New York, November 2002
  - Small molecules and proteins
  - Planned development strategy
- Denver, December 2002
  - Interactions and pathways
  - Examples: MAP kinase, Glycolysis
- San Francisco, February 2003
  - Began OWL ontology, XML Schema development

# Small Molecules

- Building blocks of pathways
- BioPAX group spent considerable time evaluating Chemical Markup Language (CML)
  - <http://www.xml-cml.org/>
- CML 2 captures desired attributes of small molecules
- Proof of concept: Successfully delivered a set of small molecules from EcoCyc to a compound visualization tool used by the Shah lab via CML 2 without loss of information
- BioPAX will use CML 2 to represent small molecules

# BioPAX Framework

A BioPAX Record has:

1. Optional set of Parts

Part has Type, Attribute(s), State(s)

2. Optional set of Interactions

Interactions have Type and Attribute(s)

3. Optional Attribute(s), including a Type

# BioPAX Framework

- Parts

- Type
  - Biological Sequence (PSI / BioPAX)
  - Small Molecule (CML)
  - Cellular Component (Gene Ontology)
  - BioPAX Record
- Attributes (Type Dependent)
- State
  - Informational or Physical

- Interactions

- Type
  - Undirected (a set of Parts)
    - » Subtype: E.g. molecular association, co-expression, co-occurrence
  - Directed (a set of inputs and a set of outputs)
    - » Subtype: E.g. biochemical reaction, molecular assembly, transport
- Attributes (Subtype dependent)

- Attributes

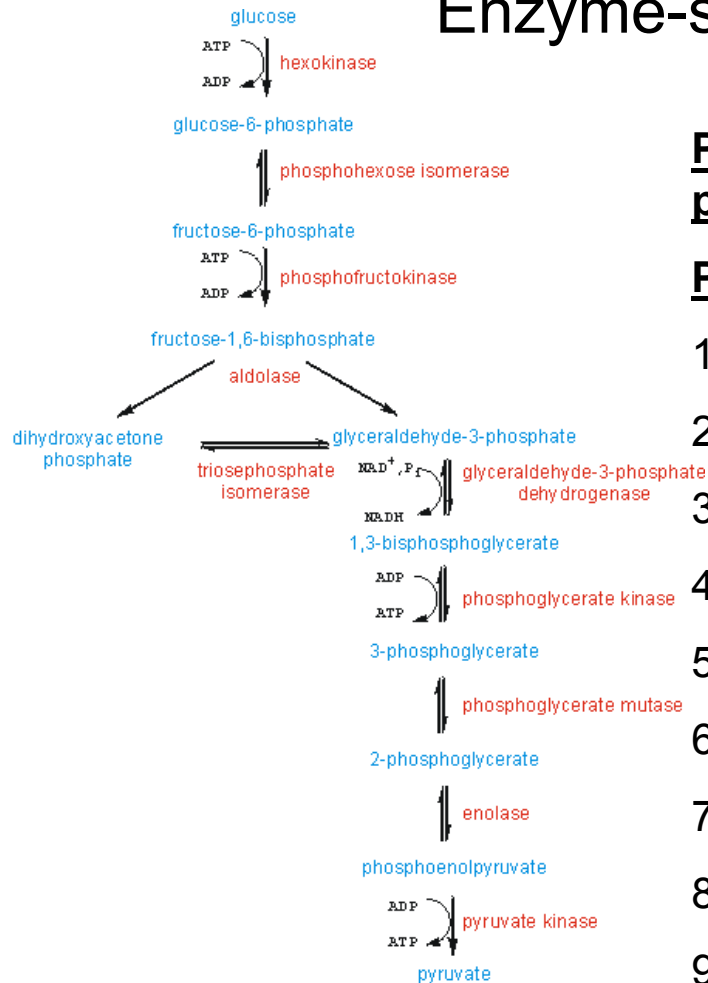
- UGI?
- Type
  - Pathway
  - Complex
  - Co-occurrence
  - Co-expression
- Name
- Timing (of Parts and Interactions)
- Evidence
  - Expt. Description
  - Expt. Conditions
  - Publication Ref.
  - Database Ref.
  - Confidence
  - Quality

# Framework → Syntax

- Currently translating BioPAX conceptual framework 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 versions will be compatible with and fully translatable to each other
- Rationale: BioPAX *must* be widely accepted to be useful, dual syntaxes will facilitate this

# Glycolysis 1

## Enzyme-substrate-product representation 1



### PAX Record: biological process

#### Parts:

1. Glucose
2. ATP
3. ADP
4. Glucose-6-phosphate
5. Hexokinase
6. Phosphohexose isomerase
7. Fructose-6-phosphate
8. Phosphofructokinase
9. Fructose-1,6-bisphosphate
- etc.

### Interactions:

1. Directed: Hexokinase (enzyme), Glucose, ATP → Hexokinase (enzyme), Glucose-6-phosphate, ADP
2. Directed: Phosphohexose isomerase (enzyme), Glucose-6-phosphate → Phosphohexose isomerase (enzyme), Fructose-6-phosphate
3. etc.

Timing:

Interaction order: 1,2,3,etc.

# Glycolysis 2

## Enzyme-substrate-product representation – nested PAX records

### **PAX Record 1: biochemical reaction**

#### **Parts:**

1. Glucose
2. ATP
3. ADP
4. Glucose-6-phosphate

#### **Interactions:**

1. Directed: Glucose, ATP → Glucose-6-phosphate, ADP

### **PAX Record 2: Catalysis**

#### **Parts:**

1. Hexokinase
2. PAX Record 1

#### **Interactions:**

1. Directed: (subtype catalysis) Hexokinase → PAX Record 1

**PAX Record 3..20 would describe more of glycolysis in this way**

### **PAX Record 20: biological process**

1. PAX Record 2
2. PAX Record 4
3. etc.

**(All of the catalysis PAX records are collected)**

#### **Interactions:**

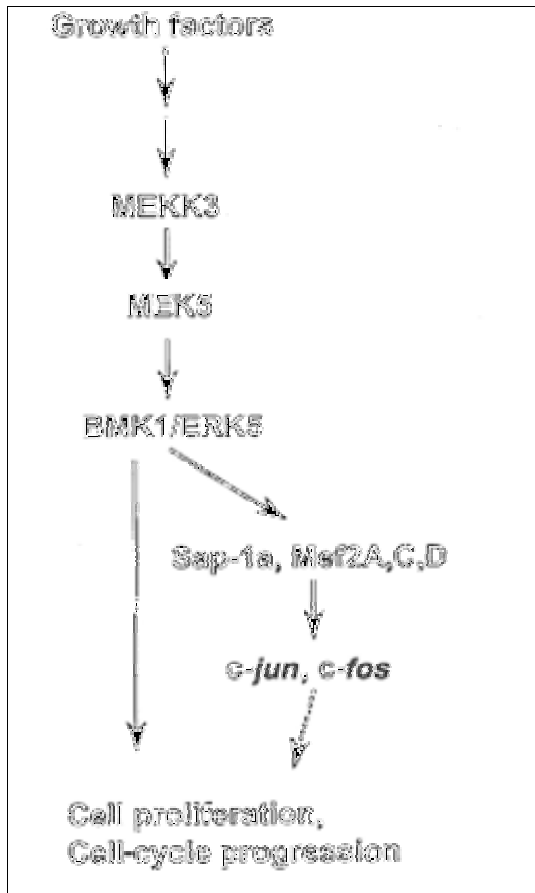
None defined.

#### **Timing:**

Part 1,2,3, etc.

# MAP Kinase Pathway 1

Signaling diagram representation of this pathway



Proposed model for the BMK1 signal transduction pathway.

from <http://www.scripps.edu/research/sr2000/imm08.html>

PAX Record: biological process

Parts:

1. Growth factor (GO:0008083)
2. MEKK3
3. MEK5
4. ERK5
5. Sap-1a
6. c-jun
7. Cell proliferation (GO:0008283)

Interactions:

1. Directed: Growth factor → MEKK3
2. Directed: MEKK3 → MEK5
3. Directed: MEK5 → ERK5
4. Directed: ERK5 → Cell proliferation
5. Directed: ERK5 → Sap-1a
6. Directed: Sap-1a → c-jun
7. Directed: c-jun → Cell proliferation

Timing:

Interaction order: 1,2,3(4&(5,6,7))

# MAP Kinase Pathway 2

State transition representation of this pathway

## PAX Record: Biological process

### Parts:

1. Growth factor (GO:0008083)
2. MEKK3 (+,phosphorylated and -,unphosphorylated states)
3. MEK5 (+,phosphorylated and -,unphosphorylated states)
4. ERK5 (+,phosphorylated and -,unphosphorylated states)
5. Sap-1a (+,phosphorylated and -,unphosphorylated states)
6. c-jun (+,phosphorylated and -,unphosphorylated states)
7. Cell proliferation (GO:0008283) (+,- states)

(Note: +,- denotes active, inactive)

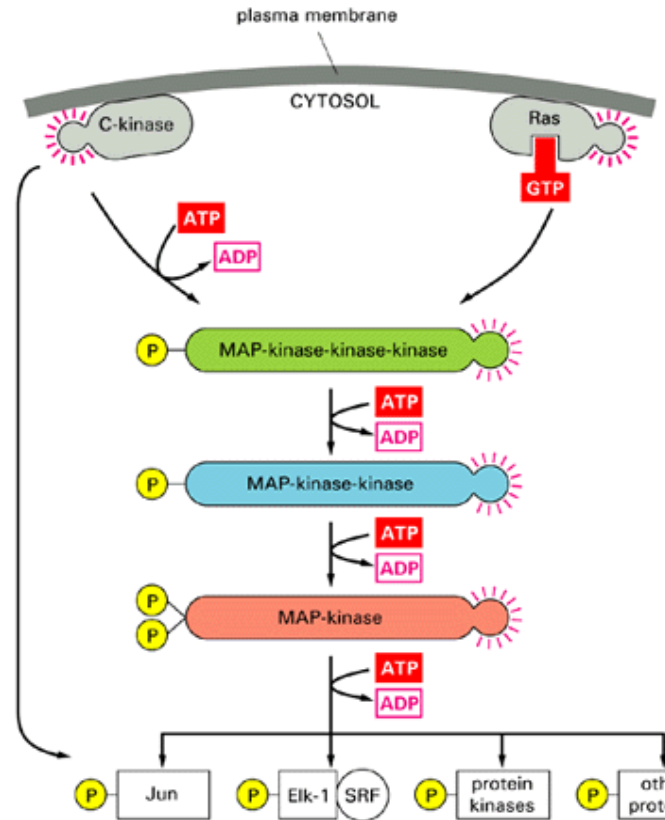
### Interactions:

1. Directed: Growth factor, MEKK3- → Growth factor MEKK3+
2. Directed: MEKK3+, MEK5- → MEKK3+, MEK5+
3. Directed: MEK5+, ERK5- → MEK5+, ERK5+
4. Directed: ERK5+, cell proliferation - → ERK5+, cell proliferation+
5. Directed: ERK5+, Sap-1a- → ERK5+, Sap-1a+
6. Directed: Sap-1a+, c-jun- → Sap-1a+, c-jun+
7. Directed: c-jun+, cell proliferation - → c-jun+, cell proliferation+

### Timing:

Interaction order: 1,2,3(4&(5,6,7))

# MAP Kinase Pathway 3



From: Molecular Biology of the Cell, 3rd edn. Part III. Internal Organization of the Cell Chapter 15. Cell Signaling

# MAP Kinase Pathway 3

Representation of this pathway with states as separate parts

## PAX Record: Biological process

### Parts:

1. Growth factor (GO:0008083)
2. MEKK3 (-)
3. MEKK3-phosphorylated (+)
4. MEK5 (-)
5. MEK5-phosphorylated (+)
6. ERK5 (-)
7. ERK5-phosphorylated (+)
8. Sap-1a (-)
9. Sap-1a-phosphorylated (+)
10. c-jun (-)
11. c-jun-phosphorylated (+)
12. ATP
13. ADP
14. Cell proliferation (GO:0008283)  
(+,- states)

(Note: +,- denotes active, inactive)

### Interactions:

1. Directed (subtype biochemical reaction): MEKK3 (-), ATP → MEKK3-phosphorylated (+), ADP
2. Directed (subtype catalysis): growth factor → Interaction 1
3. Directed (subtype biochemical reaction): MEK5 (-), ATP → MEK5-phosphorylated (+), ADP
4. Directed (subtype catalysis): MEKK3-phosphorylated (+) → Interaction 3
5. Directed (subtype biochemical reaction): ERK5 (-), ATP → ERK5-phosphorylated (+), ADP
6. Directed (subtype catalysis): MEK5-phosphorylated (+) → Interaction 5
7. Directed (subtype biochemical reaction): Sap-1a (-), ATP → Sap-1a-phosphorylated (+), ADP
8. Directed (subtype catalysis): ERK5-phosphorylated (+) → Interaction 7
9. Directed (subtype biochemical reaction): c-jun (-), ATP → c-jun-phosphorylated (+), ADP
10. Directed (subtype catalysis): Sap-1a-phosphorylated (+) → Interaction 9
11. Directed (subtype catalysis): ERK5-phosphorylated (+) → cell proliferation
12. Directed (subtype catalysis): c-jun-phosphorylated (+) → cell proliferation

### Timing:

Interaction order: (1&2),(3&4),(5&6),(11 & ((7&8),(9&10)),12)

# mRNA transcript co-expression

## **PAX Record: co-expression**

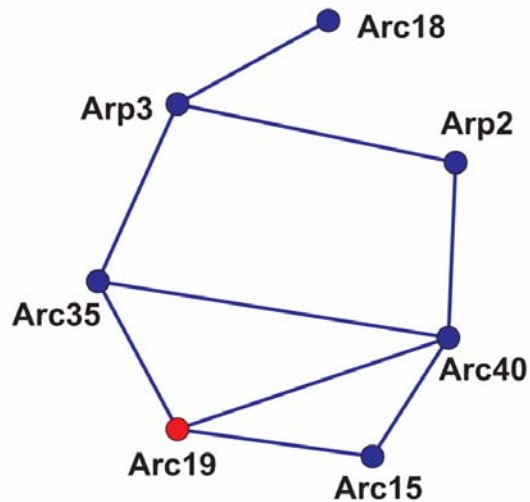
### **Parts:**

1. mRNA transcript of MEKK3
2. mRNA transcript of MEK5
3. mRNA transcript of ERK5

### **Interactions:**

1. Undirected: mRNA transcripts of MEKK3, MEK5, ERK5

# Protein-protein interaction



**PAX Record: molecular interaction**

**Parts:**

1. Arp2
2. Arp3
3. Arc15
4. Arc18
5. Arc19
6. Arc35
7. Arc40

**Interactions:**

1. Undirected: Arp2, Arp3
2. Undirected: Arp3, Arc18
3. Undirected: Arp3, Arc35
4. Undirected: Arp2, Arc40
5. Undirected: Arc35, Arc40
6. Undirected: Arc35, Arc19
7. Undirected: Arc19, Arc40
8. Undirected: Arc19, Arc15
9. Undirected: Arc40, Arc15

# Protein complex with some known topology

## **PAX Record: molecular complex**

### **Parts:**

1. Arp2
2. Arp3
3. Arc15
4. Arc18
5. Arc19
6. Arc35
7. Arc40

### **Interactions:**

1. Undirected: Arp2, Arp3, Arc15, Arc18, Arc19, Arc35, Arc40 (This annotates the complex)
2. Undirected: Arp2, Arp3 (This annotates a known protein-protein interaction within the complex)

# Protein complex assembly

**PAX Record: biological process**

**Parts:**

1. Arp2
2. Arp3
3. Arp2-Arp3 Complex
4. Arc15
5. Arc15-Arp2-Arp3 Complex
6. etc.

**Interactions:**

1. Directed: (subtype assembly) Arp2, Arp3  
→ Arp2-Arp3 Complex
2. Directed: (subtype assembly) Arp2-Arp3  
Complex, Arc15 → Arc15-Arp2-Arp3  
Complex
3. etc.

**Timing:**

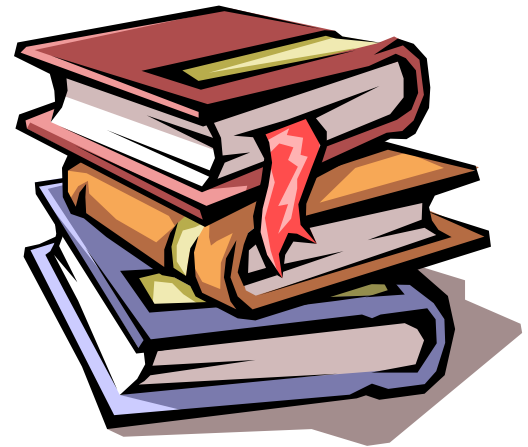
Interaction 1,2,3,etc.

# Text word co-occurrence

## BioPAX Record: Text word co-occurrence

### Parts:

1. Word glycolysis
2. Word hexokinase
3. Word diabetes



### Interactions:

1. Undirected: glycolysis, hexokinase, diabetes

# Contact BioPAX

- Participate in a subgroup
- Provide feedback
- Mailing list discussion group
- Mailing list announcements

Sign up via web site

[www.biopax.org](http://www.biopax.org)

Thank You!