

# BioPAX – Biological Pathways Exchange Language

## Level 2, Version 0.5 (Draft Release) Documentation

BioPAX Draft Recommendation, November 23, 2004

Copyright © 2004 BioPAX Workgroup. All Rights Reserved

---

### Summary

BioPAX Level 2 expands the scope of the BioPAX data exchange format to include representation of the following biological pathway concepts: black-box pathways, hierarchical pathways and molecular binding interactions. Adding coverage of these concepts will allow BioPAX to represent the bulk of the data in the PSI-MI Level 2 format (<http://psidev.sourceforge.net/mi/rel2/doc/>) and lays the groundwork for support of signal transduction and molecular states. This document describes the initial draft of BioPAX Level 2.

### Feedback

Review of this document and of the accompanying OWL file is welcome and encouraged. Comments may be sent to the [biopax-discuss@biopax.org](mailto:biopax-discuss@biopax.org) email list; subscription details available here:

<http://www.biopax.org/mailman/private/biopax-discuss/>

### Scope of this document

This document is in draft mode and only describes the features of BioPAX that are new in Level 2. It is expected that readers are familiar with BioPAX Level 1, described here:

<http://www.biopax.org/release/biopax-level1-documentation.pdf>

The final version of this document will include all documentation from Level 1 and be self-sufficient.

### Status of this document

This is the initial draft of the BioPAX Level 2 documentation. This document will be updated over time, based on community feedback. The latest version of this document can always be found here:

<http://www.biopax.org/Downloads/Level2v0.5/biopax-level2-documentation.pdf>

### Related documents

BioPAX Level 2, Version 0.5 OWL file:

<http://www.biopax.org/Downloads/Level2v0.5/biopax-level2.owl>

## Overview

As described in the Level 1 documentation, the BioPAX format is developed in levels. Each level supports the representation of an increasingly wider variety of pathway data. The focus of BioPAX Level 1 was on metabolic pathway data. BioPAX Level 2 expands the scope of the format to include support for molecular binding interactions (by implementing the PSI-MI object model in BioPAX), black-box pathways and hierarchical pathways.

In addition to representing these concepts, BioPAX Level 2 supports several new types of supplemental data, which allow more detailed annotation of pathway concepts. Specifically, Level 2 allows the representation of sequence features, such as binding sites and post-translational modifications, and description of supporting evidence for interactions and pathways.

Other changes in Level 2 include the addition of DNA as a physical entity, the addition of typed participant classes, and several new high-level utility classes that add organization to the utility class tree.

### PSI-MI

The Human Proteome Organization's (HUPO) Proteomics Standards Initiative's Molecular Interaction format (PSI-MI, <http://psidev.sourceforge.net/mi/rel2/doc/>) is designed to represent molecular binding interactions. Since this is also one of the aims of BioPAX Level 2, and since the PSI-MI format is already supported by a number of key databases, BioPAX Level 2 adds support for the PSI-MI object model. We strove to mirror the features of the PSI-MI XML Schema as much as possible in the BioPAX OWL ontology. A number of the new BioPAX classes and properties are identical to elements of PSI-MI in both name and definition. We expect that PSI-MI Level 1 and 2 data files can be converted to BioPAX Level 2 data files with minimal or no information loss.

### Backward Compatibility

Currently, any software written to work with BioPAX Level 2 will also work with data from BioPAX Level 1 because the classes and properties of BioPAX Level 1 form a subset of those in BioPAX Level 2, and none of the new concepts in Level 2 are mandatory. In other words, data that complies with BioPAX Level 1 also complies with BioPAX Level 2.

## New Features of BioPAX Level 2

This section provides a description of the additional features found in BioPAX Level 2 and the rationale for each.

### Molecular Binding Interactions

As large amounts of molecular interaction data are being produced from proteomics experiments in a large-scale way, early BioPAX development discussions identified the importance of representing molecular interaction data and it was prioritized for Level 2 inclusion.

Unlike most metabolic pathway data, which tends to represent interactions in a high level of detail, molecular interaction datasets rarely capture causal or temporal aspects of interactions. As a result, molecular binding interactions are often considered a “low resolution” form of pathway data. BioPAX Level 2 captures molecular binding interactions at a relatively high level in the ontology class hierarchy reflecting the fact that any given binding interaction may be a low-resolution, or more abstract, view of a more specific type of interaction.

For example, a signaling database would likely capture the interaction between MEK1 and ERK1 as a catalysis event (MEK1 catalyzes the phosphorylation of ERK1). A molecular interaction database, on the other hand, would likely store the interaction simply as a binary protein-protein interaction. BioPAX Level 2 supports both of these representations. Furthermore, since the catalysis class is a descendent of the physical interaction class (in which molecular binding interactions are stored), BioPAX Level 2 also implicitly captures the fact that the catalysis instance involving MEK1 and ERK1 may be a more specific representation of the corresponding physical interaction instance.

## Sequence Features

BioPAX Level 2 adopts the mechanism used by the PSI-MI format to represent sequence features. Any protein, RNA, or DNA participant may contain a sequence feature. However, it is recommended that users only define sequence features that are relevant to the interaction at hand. Separate participant instances should be created for each interaction in which different sets of sequence features are relevant.

## Evidence

Since many molecular interactions in existing databases are derived from experiments with high false-positive rates, it is important to be able to capture the experimental evidence supporting these interactions. The PSI-MI format allows detailed descriptions of experiments, which may be associated with one or more interactions. BioPAX Level 2 expands on this to allow evidence codes, such as those developed by GO or BioCyc, to be attached as evidence for interactions and pathways.

## Hierarchical Pathways

A pathway can be composed of another pathway (including itself). For example, the cell cycle is often decomposed into stages that are each considered their own pathway. To allow hierarchical pathways and recursive pathways, BioPAX Level 2 expands the pathway class to allow participation of pathways in addition to interactions.

## Black-box Pathways

It is often useful to be able to define a pathway without specifying all of the individual component interactions. The details of the pathway may not be known, for example, or they may

not be relevant to the task at hand. BioPAX Level 1 allowed empty pathways, but had no mechanism for representing the function of these pathways in terms of physical entities.

BioPAX Level 2 solves this problem by allowing pathways to have inputs and outputs. This will allow users to treat a multi-step process for which the details are unknown or not relevant as a black-box pathway. Because Level 2 also supports hierarchical pathways, black-box pathways may participate in interactions and other pathways. This will allow representation of complex cellular events in a wide range of detail.

For example, a GO Biological Process term, such as glycolysis (GO:0006096), could be used to name a pathway, without describing intermediate steps, but still allowing the pathway to be used in a network of super-pathways that link outputs of one pathway to inputs of another.

One common question regarding black-box pathways is, “What is the difference between a black-box pathway and a conversion interaction since both may convert one set of physical entities to another?” In BioPAX, we make the following distinction: a pathway has one or more known or suspected physical entity intermediates that would normally be described if the pathway were fully elucidated and translated into BioPAX format, while a conversion does not. In other words, a black-box pathway would have at least 2 pathway steps if those steps were represented in BioPAX, while a conversion would only ever have one step in BioPAX.

## Miscellaneous

### Utility class organization

With the increased number of utility classes in Level 2, a number of new organizational classes in the utility class tree were created to partition the utility class hierarchy into more manageable subdivisions.

### INChI

BioPAX Level 2 allows small molecule structures to be represented using the new IUPAC-NIST Chemical Identifier (INChI) format. This format is used to describe small molecules by NCBI's PubChem resource, among others.

## 2 Ontology Changes

This section lists all of the changes from BioPAX Level 1. Underlined property names are new properties in Level 2 and are defined at the end of this section.

### Changes to Existing Classes

#### interaction

**Change:** Added HAS-SUPPORT, NEGATIVE properties.

## pathway

**Change:** Added HAS-SUPPORT, INPUT, OUTPUT properties; added pathway to range of PATHWAY-COMPONENTS property.

## chemicalStructure

**Change:** Added “INCHI” to allowed values of STRUCTURE-FORMAT property.

## pathwayStep

**Change:** Added pathway to range of STEP-INTERACTIONS property.

## New Entity Classes

### physicalInteraction

**Parent:** interaction

**Children:** control, conversion

**Definition:** “An interaction in which at least one participant is a physical entity, e.g. a binding event. This class should be used as the default for representing molecular interactions, such as those defined by PSI-MI level 2. If sufficient information on the nature of a molecular interaction is available, a less abstract BioPAX interaction class may be used.”

**Additional properties:** INTERACTION-TYPE

### dna

**Parent:** physicalEntity

**Definition:** “A physical entity consisting of a sequence of deoxyribonucleotide monophosphates; a deoxyribonucleic acid (e.g. a chromosome, plasmid). A specific example is chromosome 7 of Homo Sapiens.”

**Additional properties:** ORGANISM, SEQUENCE

## New Utility Classes

### confidence

**Parent:** utilityClass

**Definition:** “Confidence that the containing instance actually occurs or exists *in vivo*, usually a statistical measure.”

**Additional properties:** CONFIDENCE-UNIT, CONFIDENCE-VALUE

### externalReferenceUtilityClass

**Parent:** utilityClass

**Children:** bioSource, dataSource, openControlledVocabulary, xref

**Definition:** “A utility class that acts as a pointer to an external object, such as an entry in a database or a term in a controlled vocabulary.”

## **complexParticipant**

**Parent:** physicalEntityParticipant

**Definition:** “A complex participant. This class describes any additional special characteristics of a complex, such as cellular location, that are relevant to its participation in an interaction.”

## **dnaParticipant**

**Parent:** physicalEntityParticipant

**Definition:** “A DNA participant. This class describes any additional special characteristics of a DNA molecule, such as base modifications, that are relevant to its participation in an interaction.”

**Additional properties:** SEQUENCE-FEATURE-LIST

## **proteinParticipant**

**Parent:** physicalEntityParticipant

**Definition:** “A protein participant. This class describes any additional special characteristics of a protein, such as post-translational modifications, that are relevant to its participation in an interaction.”

**Additional properties:** SEQUENCE-FEATURE-LIST

## **rnaParticipant**

**Parent:** physicalEntityParticipant

**Definition:** “An RNA participant. This class describes any additional special characteristics of an RNA molecule, such as secondary structure, that are relevant to its participation in an interaction.”

**Additional properties:** SEQUENCE-FEATURE-LIST

## **smallMoleculeParticipant**

**Parent:** physicalEntityParticipant

**Definition:** “A small molecule participant. This class describes any additional special characteristics of a small molecule, such as stoichiometric coefficient, that are relevant to its participation in an interaction.”

## **sequenceLocationUtilityClass**

**Parent:** utilityClass

**Children:** sequenceFeature, sequenceInterval, sequenceSite

**Definition:** “A utility class that describes a location on a nucleotide or amino acid sequence.”

## sequenceFeature

**Parent:** sequenceLocationUtilityClass

**Definition:** “A sequence location class that describes a feature on a sequence relevant to an interaction, such as a binding site or post-translational modification.”

**Additional properties:** FEATURE-LOCATION, FEATURE-TYPE, NAME, SHORT-NAME, SYNONYMS, XREF

## sequenceInterval

**Parent:** sequenceLocationUtilityClass

**Definition:** “A sequence location class that describes an interval on a sequence.”

**Additional properties:** SEQUENCE-INTERVAL-BEGIN, SEQUENCE-INTERVAL-END

## sequenceSite

**Parent:** sequenceLocationUtilityClass

**Definition:** “A sequence location class that describes a site on a sequence, i.e. the position of a single nucleotide or amino acid.”

**Additional properties:** POSITION-STATUS, SEQUENCE-POSITION

## support

**Parent:** utilityClass

**Children:** experiment, experimentalForm

**Definition:** “A utility class that describes the support for a particular assertion, such as the existence of an interaction or pathway.”

**Additional properties:** HAS-CONFIDENCE, HAS-EVIDENCE-CODE

## experiment

**Parent:** support

**Definition:** “A utility class that describes a single experiment. This class is used to support the existence of an interaction or pathway.”

**Additional properties:** FEATURE-DETECTION-METHOD, HOST-ORGANISM, INTERACTION-DETECTION-METHOD, NAME, PARTICIPANT-DETECTION-METHOD, XREF

## experimentalForm

**Parent:** support

**Children:** dnaExperimentalForm, proteinExperimentalForm, rnaExperimentalForm

**Definition:** “A utility class that describes the form of a physical entity in a particular experiment. This class is used to support the assertion that a particular physical entity participates in an interaction or pathway.”

**Additional properties:** EXPERIMENTAL-ROLE, IN-EXPERIMENT, PARTICIPANT

## **dnaExperimentalForm**

**Parent:** experimentalForm

**Definition:** “A utility class that describes the form of a DNA molecule in a particular experiment. This class is used to support the assertion that a particular DNA molecule participates in an interaction or pathway.”

## **proteinExperimentalForm**

**Parent:** experimentalForm

**Definition:** “A utility class that describes the form of a protein in a particular experiment. This class is used to support the assertion that a particular protein participates in an interaction or pathway.”

**Additional properties:** IS-OVEREXPRESSED, IS-TAGGED

## **rnaExperimentalForm**

**Parent:** experimentalForm

**Definition:** “A utility class that describes the form of an RNA molecule in a particular experiment. This class is used to support the assertion that a particular RNA molecule participates in an interaction or pathway.”

**Additional properties:** IS-OVEREXPRESSED

## **New Properties**

CONFIDENCE-UNIT	The unit of the confidence measure.
CONFIDENCE-VALUE	The value of the confidence measure.
EXPERIMENTAL-ROLE	The role of the participant in the experiment, e.g. "bait" or "prey".
FEATURE-DETECTION-METHOD	Method used to determine the features of the interaction participants.
FEATURE-LOCATION	Location of the feature on the sequence of the interactor. One feature may have more than one location, used e.g. for features which involve sequence positions close in the folded, three-dimensional state of a protein, but non-continuous along the sequence.
FEATURE-TYPE	Description and classification of the feature.
HAS-CONFIDENCE	Confidence in the containing instance. Usually a statistical measure.
HAS-EVIDENCE-CODE	A pointer to a term in an external controlled vocabulary, such as the GO or BioCyc evidence

	codes, that describes the nature of the support.
HAS-SUPPORT	Scientific evidence supporting the existence of the entity as described.
HOST-ORGANISM	The host organism in which the experiment has been performed.
IN-EXPERIMENT	The experiments in which the participant has the experimental form being described.
INPUT	Input to a pathway. Can be used with a 'black-box' pathway that does not include any steps.
INTERACTION-DETECTION-METHOD	Experimental method used to determine the interaction.
INTERACTION-TYPE	External controlled vocabulary characterizing the interaction type, for example "phosphorylation".
IS-OVEREXPRESSED	If true, the participant was over-expressed in the experiment.
IS-TAGGED	If true, the participant was tagged in the experiment
NEGATIVE	If true, this interaction has been shown NOT to occur. This slot is optional and if no value is present, then the interaction is assumed to occur (i.e. the default value for this slot is false).
OUTPUT	Output from a pathway. Can be used with a 'black-box' pathway that does not include any steps.
PARTICIPANT	The participant that has the experimental form being described.
PARTICIPANT-DETECTION-METHOD	Method used to determine the identity of the interaction participants.
POSITION-STATUS	The confidence status of the sequence position.
SEQUENCE-FEATURE-LIST	Sequence features relevant for the interaction, for example binding domains or post-translational modification sites.
SEQUENCE-INTERVAL-BEGIN	The begin position of a sequence interval.
SEQUENCE-INTERVAL-END	The end position of a sequence interval.
SEQUENCE-POSITION	The integer position gives the position. The first base or amino acid is position 1. In combination with the numeric value, the property 'POSITION-STATUS' allows to express fuzzy positions, e.g. 'less than 4'.