

BioPAX Work Group January 2005 Meeting in NYC

The January 2005 BioPAX meeting focused on resolving open Level 2 issues and on planning the development of Level 3. Proposals for Level 2 issues were discussed and refined during a pre-meeting workshop. These were presented during the first half of the main meeting; during the second half participants discussed possible features of Level 3.

In both the pre-meeting and the main meeting, short presentations were given to introduce each topic and present the major issues. Open discussions followed, in which participants offered feedback on the ideas presented. This document provides links to the presentation slides and lists the major points made during these discussions.

Dates:

January 20, 2005 – Pre-meeting workshop

January 21, 2005 – Main meeting

Location: Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 307 East 63rd St, New York, NY.

Participants:

Name	Organization	Project
Mirit Aladjem	National Cancer Institute	eMIM
Gary Bader	MSKCC	BioPAX
Michael Cary	MSKCC	BioPAX
Kam Dahlquist	Vassar College	GenMapp
Emek Demir	Bilkent University	PATIKA
Peter D'Eustachio	New York University	Reactome
Ken Fukuda	CBRC	INOH
Frank Gibbons	Harvard University	BioGraphNet
Marc Gillespie	Cold Spring Harbor Lab	Reactome
Michael Hucka	Caltech	SBML
David Kane	National Cancer Institute	eMIM
Christian Lemer	Université Libre de Bruxelles	aMAZE
Joanne Luciano	Biopathways Consortium	BioPAX
Suzanne Paley	SRI	BioCyc
Elgar Pichler	AstraZeneca	OntoSieve
Jonathan Rees	Millennium	PARIS
Andrey Rzhetsky	Columbia University	GeneWays
Chris Sander	MSKCC	BioPAX
Andrea Splendiani	Università di Milano Bicocca	GCA
Mustafa Syed	Argonne National Lab	WIT/PUMA2
Edgar Wingender	BIOBASE	BIOBASE
Guanming Wu	Cold Spring Harbor Lab	Reactome
Jeremy Zucker	Dana Farber	BioPAX

January 20, 2005 Pre-Meeting Notes:

1. [Overview of Level 1](#)

- a. Physical entity participants (PEPs) discussion
 - i. Allow different interactions to use the same PEPs?
 1. Currently: Yes
 2. Danger: Unwitting side-effects, e.g. updating a PEP to provide additional detail for a specific interaction might invalidate it for other interactions
 3. Recommendation: Unique PEPs for every interaction
 - ii. Make multiple copies of PEP instead of using stoichiometry slot?
 1. Would require separate instances for each PEP copy
 2. Recommendation: Use stoichiometry slot instead of making multiple copies of PEPs.
 - b. Representing polymerization: This is an open question (for a later level)
 - c. Complexes
 - i. Allow building from other complexes?
 1. Currently: Yes
 2. Danger: Same complexes can be represented multiple ways
 3. Recommendation: Do not use complexes in components slot. Complexes should always be represented flatly, i.e. not hierarchically
 - ii. What about black-box complexes (no components listed)
 1. E.g. suppose you want to say ribosome is made of small and large sub-units, but don't want to specify protein and RNA composition
 2. Recommendation: Need to allow hierarchical construction of complexes, but only when using black-box complexes.
- ## 2. [Level 2 Introduction: Representing Molecular Interactions](#)
- a. Genes vs. DNA
 - i. Need DNA for protein:DNA binding (level 2)
 - ii. Need genes for gene regulation, genetic interactions (later level)
 - iii. Later, we will add the Gene class. DNA unification xrefs point to DNA sequence databases, like GenBank. Gene unification xrefs will point to gene databases, like Entrez Gene and not to DNA sequence databases.
 - b. Direct / Indirect
 - i. Allow stating whether and interaction is direct or indirect?
 1. Currently: No
 2. Recommendation: Boolean slot on interaction (Direct: yes/no) (done)
- ## 3. [Level 2 Issue: Experimental Evidence](#)
- a. Controlled Vocabularies (CVs)
 - i. Not standardizing CVs (e.g. for evidence code) makes integration difficult (hard for importers to deal with all CVs)
 - ii. Mission-critical CVs (e.g. localization) need to be more controlled
 - iii. Recommendation: Require (or strongly recommend) using specific CVs for mission-critical slots, e.g.:

1. Localization (GO cellular component)
 2. Taxonomy (NCBI taxonomy)
 3. Evidence code (combo of BioCyc and PSI-MI evidence CVs)
- iv. Mirror these recommended CVs on biopax.org + package with distribution
 - v. Some CV terms are “icing”, extra non-critical annotation (practically free-text): Allow these to use whatever CVs users want (e.g. cell type in bioSource, interaction type)
 - vi. Open issue: PSI-MI CVs are protein-centric, need to make more generic / inclusive
 - vii. Open issue: what other CVs are available for use e.g. iNOH, WIT – mention these in the documentation.
 - viii. Open issue: versioning for CVs terms – the xref.version property is available for this, but we need to better define its use.
- b. Raw data vs. “believed” interactions
 - i. Should we distinguish between experimental observations (e.g. an interaction observed once in a Y2H experiment) and the interactions inferred from this evidence?
 - ii. No, at least not in Level 2. Interactions may be supported by multiple pieces of evidence, so allow users to determine truth on their own based on stated evidence. Also, “raw” data is often full of interpretation.
 - c. Other issues
 - i. Negative information
 1. E.g. “this interaction is known not to occur”
 2. Issue: may be error-prone if user accidentally thinks negative is positive (fails to check the ‘negative’ property)
 3. Few DBs capture this, put it off until later levels
 - ii. Open issue: Using one interaction as evidence for another (e.g. a protein-protein interaction as evidence for a biochemical reaction)
 - iii. Experiment class is highly specific to protein-protein interactions - make it more generic (and simpler!) (done)
 - iv. First version of recommended CV for evidence code should be integration of BioCyc and PSI terms
 - v. Make a requirements document
 - vi. Host organism points to biosource – what about ‘in vitro’?
 - vii. Confidence unit – use a CV? – No
 - viii. Create a general mechanism for provenance, evidence, source DB? This can be added in later levels if there is need. Currently just have data source – last source of this data.
 - ix. Need indication that protein is not in natural state in experiment (e.g. tagged, ectopic expression, from foreign species, etc.) – experimental form in level 2 handles this.
 - x. Support (evidence) for complexes? (Evidence only allowed on interactions currently)
 1. Could move evidence up to entity class

- 2. Recommendation: not in level 2
 - xi. Binding site: is it entire binding site or just a part of it – make this explicit in docs.
- 4. [Level 2 Issue: Sequence Features](#)
 - a. Disulfide bridges?
 - i. Should we make sequence features entities which may participate in interactions? (could be used to describe complex topology explicitly)
 - ii. Argument against: complicates queries
 - iii. Recommendation: not in level 2
 - b. Remove sequence features from sequenceLocation utility class (done)
 - c. Constant features (i.e. annotations) vs. modifications (e.g. PTMs)
 - i. Put constant features directly in entities (e.g. protein class)?
 - ii. Partition sequenceFeatures into two subclasses?
 - 1. E.g. Constant & Variable?
 - iii. Recommendation: wait for state subgroup to decide
 - d. Sequence feature representation is related to issue of states – re-form state subgroup to address

January 21, 2005 Main Meeting Notes:

1. Introduction
 - a. [Participant introduction slides](#)
 - b. [Survey results](#)
2. Review of Level 2 and pre-meeting discussions
 - a. Recommend specific CVs for certain slots
 - i. Mirror these on biopax.org
 - b. Rework evidence and confidence
 - c. Add DIRECT (yes/no) slot to physicalInteraction (done)
 - d. Point from experiment to expForm (not vice versa) (done)
 - e. IS_OVEREXPRESSED, IS_TAGGED → Separate ontology (CV)
 - i. No, too important (compare to EC numbers, which we have slots for even though they are XREFs)
 - ii. Post-meeting discussion: Expand EXPERIMENTAL-ROLE slot to take all external CV descriptors (e.g. His-tagged, TAP-tagged, bait, prey) of experimental forms
 - f. Requested: Protein fragments as a possible experimental form of proteins
 - g. Sequence features
 - i. Open issue: should feature type be mission-critical CV?
 - h. Unique PEPs for each interaction (1.A.i from pre-meeting)
 - i. Issue: How to link one interaction to next in pathway if no common PEPs?
 - ii. After discussion, decided it's not an issue, because explicit linking is arbitrary. State representation will make this easier.
 - i. Desired documentation: Concrete use cases and requirements for each feature
 - j. KEQ needs to specify which direction (e.g. LEFT→RIGHT) it applies to
3. [Level 3 Issue: Hierarchical Pathways](#) – will be included in level 3

4. [Level 3 Issue: Pathway Input and Output](#)
 - a. Desired: mechanism of composition (as alternative to subclasses with dual-inheritance) – Yes, will be included in level 3, representation not yet decided.
5. [Level 3 Issue: State Representation](#) – re-form state subgroup to finalize state representation.
6. [Level 3 Issue: Gene Regulation](#)
 - a. Do we want promoters and other regulatory regions modeled in BioPAX?
Decision: not for level 3, maybe later.
7. [Potential Level 4 Issue: Generic entities](#)
 - a. Polymerization should be added to the list of generic interactions
 - b. Decision: yes, deal with in level 4
8. [Potential Level 4 Issue: Genetic interactions](#)
9. Development Timeline
 - a. Alpha → Beta → Release candidate → Release
 - i. Level 2 – Release by June 25
 - ii. Level 3 – Beta by July 25
10. Desired Software
 - a. Validation suite (part of libBioPAX)
 - b. Editor & visualizer (initially Protégé tools are available)
11. Documentation
 - a. Level 2 docs by March 15 (beta stage)
 - b. Use cases (post on biopax.org)
 - c. Requirements doc
12. Level 3 Subgroups (a way of organizing interested participants and communicating to the outside world what our needs are and who should be contacted with specific requests)
 - a. States (contact: Emek Demir)
 - b. External vocabularies (contact: Jeremy Zucker, John Rees)
 - c. Abstractions (e.g. hierarchical pathways) (contact: Ken Fukuda)
13. Other possible subgroups
 - a. Funding/support/sponsorship (contact: Mike Cary, Gary Bader, Joanne Luciano)
 - b. Documentation (contact: Mike Cary, Gary Bader)
 - c. Tools development (contact Gary Bader, Joanne Luciano, Jeremy Zucker)
 - d. We should post these on the how to participate section of the website.

Post meeting comments: Christian Lemer

-Best Practice (BP): Put all names in synonyms i.e. include the short label

-Explain the organism slot on pathway better in docs

BP: URL in publicationXref – don't use unless document is not in PubMed

Question: describe PTM of an unknown site?

Docs: be more explicit about data source – it's only last data source, not full provenance.

If you want full provenance, you must get it from the database you got the data from.