

BioPAX Working Group
September 17, 2003 Conference Call Minutes

Participants: Gary Bader, Mike Cary, Chris Sander, Peter Karp, Joanne Luciano (wrote minutes), Imran Shah

Agenda

1. Ontology work/status
 - a. Version 0.5 release
 - i. Goal: Have it out by Friday
 - ii. General comments on release process?
 - iii. Should we submit specific questions to the discuss list?
 1. Try to focus discussion around specific issues?
 - a. If so, which issues?
 - iv. Mechanism for receipt of comments/bug reports
2. External meetings
 - a. Report: GK conference call
 - b. Report: OMG meeting in Boston
 - c. Are there any upcoming meetings / conferences that should be on our radar screen?
3. Subgroup status reports
 - a. State subgroup
 - b. Small molecule subgroup
 - i. Erik Brauner unable to attend 10/3 meeting
 - c. Examples subgroup
 - i. How should we track / log examples?
 - d. XML Schema subgroup
4. Next conference call date
 - a. Do we need one? If so, October 1?
5. Next F2F meeting: Oct.3 in New York
 - a. Administrivia
 - i. Travel, hotel, dinner arrangements
 - b. Discuss agenda (see http://www.biopax.org/pm/pmdocs/10-03-03_f2f-agenda.htm)
 - i. Ontology issues
 1. Context
 2. Equivalence class
 3. Central dogma
 4. Provenance
 - c. Second session: Saturday Oct. 4
 - i. Who plans to attend?
 - ii. Work through some examples?

Summary

- Scheduled release of version 0.5 released this Friday, September 19, 2003 looks doable.
- Comments and bug reports should be submitted to the BioPAX-discuss list, which will be tracked in Sourceforge.
- Need issue specific examples and instances from user community
- BioPAX met with the following external groups:
 - The Genome KnowledgeBase (GK) project is a collaborative effort (CSHL, EBI, GO) to develop a curated resource of core pathways and reactions in human biology. The meeting was a conference call September 17, 2003. Initial discussion and introduction of GK and BioPAX to each other.
 - OMG Meeting in Boston, MA, USA – Pathways Session September 10, 2003, Presented BioPAX background, history, role and relation to pathway efforts especially SBML and CellML.
 - W3C Meeting in Cambridge MA, USA – Evaluating early forms of BioPAX within an RDF Semantic Web graph model.
- **Subgroup reports**
 - **States Subgroup**
 - Received good comments from Aviv and Vincent
 - Mike and Gary will integrate feedback into the ontology
 - Getting more interest from the community which will facilitate review
 - Close to completion and includes an executive summary
 - **Examples Subgroup**
 - A message was sent to the mailing list to expect the Version 0.5 release.
 - **XML Schema Subgroup**
 - Half done and hope to be finished this week. It will accompany the Version 0.5 release.
- **Next Conference call**
 - Cancelled – not needed two days before the face-to-face meeting in NYC. Instead email will be used for reminders, issues, and announcements.
- **Review Agenda for Next Face-to-Face Meeting in New York City October 3rd, 2003**
 - Review the release v0.5 ontology
 - Review state subgroup recommendations
 - Review issues logged into Sourceforge

Detailed Notes from Conference Call

Ontology Work Status:

Update/Status report by Mike:

We want to release v0.5 this Friday Sept 19. There's not too much left to do. The remaining tasks are

- Update the way we represent states in GKB and mirrored in Protégé
- Add instances to the ontology (most significant amount of work remaining)
- Complete the documentation (A significant amount done, some remains)

The Conclusion is that we hope to get it done by Friday and that we have a good group of interested parties to provide feedback about the ontology.

Mike asked if there were any comments on the general release process noting that we talked previously about what this release would consist of.

Chris asked if the release mechanism is to post it on the web site and if it is mostly for internal release or are we releasing to a selected audience.

Mike said that we would be releasing it to the BioPAX discuss list. Chris asked the size, Gary looked it up - it is currently 87.

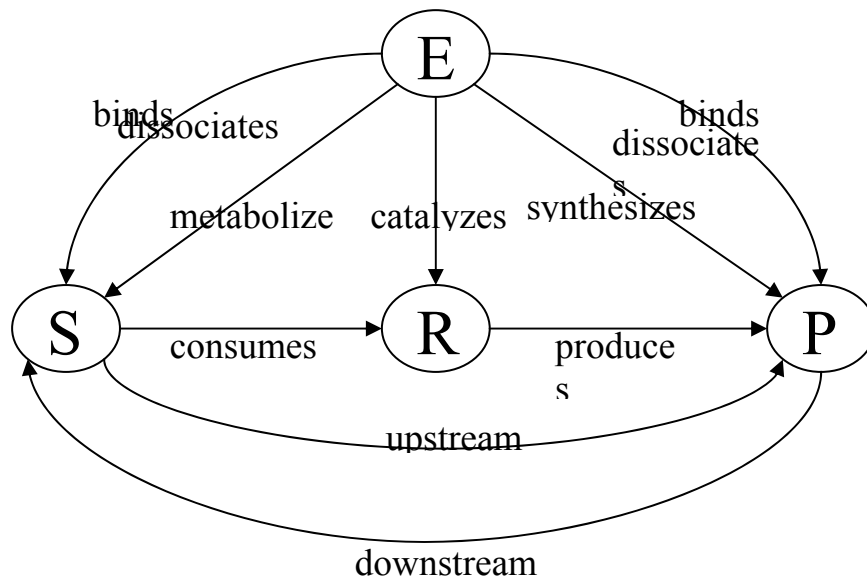
Joanne said she sent a message to the example mailing list to expect to see Version 0.5 very soon. The example's subgroup is very interested and awaiting this release.

Mike summarized that we have a significant body of people and we expect to get good feedback on our release from them.

Ontology Issue

That reminded Joanne of an issue Eric Neumann asked her to bring to the group: the BiochemReaction -> EnzymeReaction issue. Eric said he didn't like the name control (unfortunately, Eric was not present for the call where this was discussed in detail). Eric's proposal is that enzymatic reactions should have two parents, one being "biochemical reaction" and the other "control." (Unless there's a case where there's an enzymatic reaction that doesn't have an enzyme that catalyzes it).

A diagram Eric drew for Joanne:



Organization of the feedback from BioPAX release 0.5

Mike asked the group if we want to organize it in such a way that we pose specific questions about the ontology to the biopax-discuss list. That is, do we want to focus the questions around specific aspects of the ontology?

Joanne suggested that it would be useful to start setting up a series of questions and test cases that would become a benchmark or validation suite for this process. For example, known specific cases that need to be covered making sure they can be exchanged without data loss, as we did with CML. We would want to keep the set around and add to it so that with every new release, we could do what is in the software industry called regression testing. So when new features (classes) are added, old representations don't break. We would need to keep track of the various examples, use cases, and test cases so that we could validate new releases.

Should we ask people to create instances (using protégé if they have it)? We could ask the user community to come up with the test cases and how to test them out. We could also ask the group for help in putting the suite together. Gary said he thinks it would be great if they could put examples into protégé or GKB and communicate on the biopax-examples email list to give their examples so everyone can use them.

Bug reporting

Should every bug be put on the discuss list? How will the bug reporting be organized? How will significant bugs be differentiated from minor bugs?

For example, some bugs will be of the form "there should be another field for x" or "this shouldn't be an integer it should be a string" which is very different from problems with representation in the ontology.

We discussed setting up a separate email address for bug reporting and decided not to create another mailing list because it would make things too complex, however, we agreed that we need to track issues. The mechanism we will employ will be to report all bugs on the discussion list and filter the discussion list entering the bugs into a tracker on Sourceforge. The bug tracker on Sourceforge allows you to distinguish minor bugs from major issues. There are a lot of fields that can be defined with respect to bugs.

Joanne asked Gary if an interface could be set up so that someone could email a bug report and it would get automatically recorded in Sourceforge without the user having to go to the site and enter it. Gary said it is possible to set up that kind of thing and there is a system called RT that allows you to do that, but we haven't done that. Joanne said her concern is accessibility and ease for the users. The Sourceforge site isn't the most accessible or easy to use. When someone finds a bug, we want to make it easy for them to report it, we don't want them to have to jump through hoops to report it. Gary agreed and said his point was that they can just report it on the discuss list and he and Mike would moderate it and post it in Sourceforge. He and Mike are currently the only ones on the Sourceforge list and anyone else who wants to be on the list needs to create an account. Joanne asked if now, when something gets sent to BioPAX discuss, does everything

get moderated before it's posted. Gary said No. Joanne asked if it would be changing, she is concerned about flooding that list. Gary said if he thought everything being reported to the discuss list is fine. He said he didn't think it would get too crazy and if it does, then we'd do something about it. Gary said the reason he didn't think it would get too crazy is because he's on the SBML list and it's not too crazy.

Conclusion: The mechanism for reporting bugs will be the BioPAX discuss list, all bugs from small to large will be submitted there.

External Meetings

GK Group: Conference call this morning (September 19, 2003) Gary, Imre Vastrik, Eric Neumann

Background: The Genome KnowledgeBase (GK) project is a collaboration among Cold Spring Harbor Laboratory, The European Bioinformatics Institute, and the Gene Ontology Consortium to develop a curated resource of core pathways and reactions in human biology. The information in this database is authored by biological researchers with expertise in their field, maintained by the GK editorial staff, and cross-referenced with the sequence databases Ensembl and SwissProt.¹

Eric Neumann (who met Ewan Birney at the Gordon conference recently) coordinated with GK (Genome Knowledge base genomeknowledge.org). It is a database of pathways done through Ewan Birney and Lincoln Stein's groups. Eric Neumann was able to organize the call because he met Ewan Birney at a meeting recently (Gordon Conference). The call was with the main developer of GK, Imre Vastrik and the main curator, Peter D'Eustachio. They have an interesting database. Gary wrote up 4 pages of notes which he will send to the list. Basically GK is storing human biology in a very detailed way, only using the representations of enzymes, substrates and products although they have some extra things in their ontology for representing black boxes and generic things that aren't very specific, such as a generic glycolysis pathway. They don't deal with context information, like tissues, everything is talked about in a generic cell. They deal with states as unique entities, so a protein and its phosphorylated version are represented as two different entities. Also, a protein in the cytoplasm and a protein in the nucleus are represented as two different entities in their system. They have five curators working for them. They have a few hundred SWIS-PROT proteins in GK and some manner. They have a full ontology and they are using Protégé to make their ontology. Their ontology is free, it's open source and it can be easily downloaded. You can download all their data from their web site. One way of entering data, is curating from existing literature with their curators (not so interesting). The interesting way is they have a mechanism where they hook into the Cold Spring Harbor meetings to get data entered. They identify active investigators who are interested in entering data at a meeting. Working together, the biologists and curators divide up the domain and outline the general events in a pathway using a series of powerpoint cartoons that they have. They then enter the data on the PowerPoint forms, and the curators go through the results of these jamborees. To polish the material, they enter it into a version of GK for review by the authors, who review it before it goes public. That's been going on for 3 or four months. Gary's not sure how much data they have.

¹ <http://www.genomeknowledge.org>

Chris asked who at CSH is overseeing the process. Gary said there's a team of 5 curators at CSH. They sit there with the people at the meeting. It's not clear if they are full or part time. Gary said he will be meeting with them and will find out more. He added that they don't want to capture the latest experimental results, and don't care about yeast-2-hybrid results or the 'latest paper', instead, they aim to be between the Albert's textbook², and the leading journal, closer to the text book side. They are going on the detailed biochemical level, but only for human data. The CSH people are in NY at NYU. They're not full time. One person they spoke to, Peter D'Eustachio is a professor at NYU. They will meet with him in a couple of weeks after he sees Version 0.5 and gets to digest it a bit then find out more about the project. It seems like an interesting project. They are entering data, they are committed to it, and so it should be a good data source for people. They might also have some interesting aspects of their ontology that we would check against BioPAX to make sure we're covering it.

Chris brought up their background, asking if it was The Cell Migration Consortium that funded them. Gary said they didn't talk about that, Chris said he saw that on their web site and that the project was set up a while back. The consortium gave Lincoln Stein and Bill \$400,000 for 2 years to aggregate the content.

Summary: Gary will send the notes out to the group.

Mike asked if anyone had a chance to look at the GK ontology. Joanne and Mike had looked at it briefly. Chris had looked at the website but not the ontology. He noted there is a DB model and a schema. Gary noted that there are only a few new things that they have in there that aren't part of a normal enzyme-substrate-product type of reactions and pathways, pathways are collections of reactions.

Chris said it would be useful to have some understanding of where they are coming from. He pointed out, for example, in the physical entity they have concrete entity, generic entity, and simple entity. In events it is similar - they have generic events (pathways and reactions), and concrete events. Gary pointed out that that is, indeed, the main difference, i.e. generic versus concrete would be a real reaction and generic would be, for example a generic glycolysis reaction or part of glycolysis where there are a number of isozymes³ that can catalyze the same reaction.

Joanne asked if it were the case if we would work towards a common ontology with GK or we would be building a tool to map their nomenclature to ours. Gary said we don't know yet. They didn't get to that level of discussion yet, but it would be a great thing to talk about. Joanne said it would be ideal if we could come to agreement on a common ontology or we need to have good tools for mapping [JSL: we'll need good tools for mapping anyway and will start a subgroup if necessary]. Peter commented that isn't that what everyone else is doing - we're all figuring out how to map it to BioPAX. Joanne said, that's part of it, but ideally it's nice to converge on a common language. Gary said it might not be possible all the time although it would be nice. Mike said he wants to know whether we can represent the generic and concrete identifiers they

² *Molecular Biology of the Cell*, Bruce Alberts, Julian Lewis, Martin Raff, Alexander Johnson, Keith Roberts, May 2002 ISBN: 0815332181.

³ Isozyme, isoenzyme: any of two or more chemically distinct but functionally similar enzymes

have throughout their ontology and that we need to answer that. Joanne clarified that for some of the newer databases, for example Sylvia Nagl's, when she was there, there were definitely changes they were making to use BioPAX. Gary said this is good because there were some mapping issues. Joanne said she had reviewed their model and indicated where changes needed to be made to have a consistent representation. Joanne said, yes, for existing databases it's harder to come up with a common language, but for newer ones, it is easier and that both will exist in practice.

OMG Meeting in Boston (report by Joanne, who attended the OMG meeting)

This meeting was similar to the I3C meeting in which many attendees were unfamiliar with BioPAX, so Joanne gave an overview of the BioPAX initiative - its background, current status, vision and participants. OMG had put out a request for a proposal for pathway representation. Joanne reported that two proposals were received by OMG. One we knew about - the proposal submitted by Lion which proposes that SBML and CellML be adopted as the standard. A second proposal, which was from the Centre National de la Recherche Scientifique (CNRS), and proposes a UML model, SB-UML, be adopted as the standard. The first proposal was not substantial in that it was a reference to SBML and CellML and mentions BioPAX, but was not a real proposal document that could be reviewed, nor was anyone from Lion there to present it. The proposal itself was more of a reference document pointing to SBML and CellML as the proposal. It was not a self-contained proposal. It was 9 pages.

The second proposal was presented by Magali Roux from the Centre National de la Recherche Scientifique (CNRS) DAE-SDV. The title of the 40 page proposal is "SB-UML Integrating Data on Biochemical Pathways." It is a UML model for representing pathways. Joanne has a paper copy of the group, and will get an electronic copy to forward (with permission) to the PAX group.

The reviews of the proposals will be conducted offline because full proposals need to be submitted.

To check whether there are any concepts represented in the SB-UML model that we are not representing in BioPAX. Mike asked if SB-UML was based on SBML. Joanne said no, this is an independent effort by biologists in CNRS-DAE-SDV Biology Health Technology, in Paris, France. Magali is interested in looking at the BioPAX representation also.

Magali knows Vincent Schachter. Mike asked if SB-UML is being used by anyone. It is Joanne's understanding that they are using it. They developed it, are using it, and proposed it to OMG in response to OMG's RFP. It is a full proposal, with only a glossary missing which was an oversight and would be added. The SB-UML proposal is 40 pages. There were also two men from the Japan Biological Informatics Consortium who also were just hearing about BioPAX at that meeting. Joanne added them to the recipients of the email she sent out to the example's list alerting people that a release of BioPAX would be forthcoming. Another person in attendance was Martin Senger from EBI, who works on myGrid. myGrid is a multi-organizational project funded by the EPSRC as part of the UK Research Councils e-Science program and aims to develop the necessary middleware (e.g. provenance, service discovery, workflow enactment,

change notification & personalization) that will operate over an existing Web services & Grid infrastructure to support scientists in making use of complex distributed resources. For example, myGrid should enable an "e-Biologist's" workbench. myGrid utilizes standards & technologies developed for the Internet, Grid, Web services & the Semantic Web. Manchester is also involved in it (Carole Goble), and David Benton from GSK. Joanne's conclusion is that it is good to attend these meetings if they are convenient because other pathway interested parties are present so we can learn about the larger user community and they can learn about BioPAX. Others agreed that it's a good idea.

W3C Meeting at MIT 16 September 2003

Joanne was invited by Eric Neumann to join a meeting with the W3C evaluating early forms of BioPAX within an RDF Semantic Web graph model, which would allow clustering of knowledge defined in database graph standards internally. Additionally, it supports an annotation model that is machine readable, which is the main thrust of the semantic web effort and of great value to the life science community.

Attendees: Marja Koivunen (marja@w3.org), Brian Gillman from the Whitehead Institute (OmniGene), Eric Neumann, and Ralph Swick (W3C).

Gary asked "How do they fit in the OWL movement which is also in W3C?" We didn't have an answer; it wasn't talked about at the meeting.

Subgroup report

States Subgroup

Mike reported that Aviv sent a good deal of comments to the state sub group and so did Vincent Schachter. Mike said he and Gary would be meeting later today to integrate everything back into the state document. Mike said that this morning Emek Demir from the Patika group contacted him and said that they are interested in working with the states group. Mike said when he and Gary are finished with updating the document he'd send it to Emek for review. Gary added some background – they are still cycling on the states document which lay dormant for a couple of periods but it is now nicely back on track and very nearing completion. It's about 10 pages and very soon would be sent out to the pax list as a draft so they can get everyone else. They tried to structure it so that all the details are at the end of the document, which is in the style of an 'executive summary.'

Small Molecule Subgroup

Mike reported that Eric Brauner was contacted and invited to the next face-to-face meeting, but would not be able to attend. Perhaps he could attend the next meeting.

Examples Subgroup

Joanne sent a message to the examples mailing list to let them know to expect the version 0.5 release. She copied this to the BioPAX group and sent a copy to other users who would be

interested in knowing about the release such as Sylvia Nagl's group, Alan Rector (OWL/Protégé), the attendees from the OMG meeting.

XML Schema Subgroup

Gary reported that it's half done and hopefully would be finished this week. It will accompany the Version 0.5 release.

Upcoming meetings, calls

It was decided to not have the next scheduled conference call which would be two days before the face-to-face meeting and instead have an email reminder. We will take that time to check email and perhaps have some last minute discussion.

Chris is interested in finding a way to do the conference calls with a less expensive service, perhaps a net conferencing service. Joanne said she'd find out the name of a service a colleague uses.

The next meeting will be October 3rd in NY. Mike announced we would be using the money we received from DOE to fund hotel and travel expenses for people coming. Mike will send Rita's email address. Rita will make the hotel reservations. Mike encouraged everyone to send her an email as soon as possible so the hotel doesn't fill up. You will need to make your own travel arrangements and then submit them for reimbursement. Please schedule them at least two weeks in advance to get better fares. Thursday, the night before, we plan to go out to dinner. Imran and Peter will be coming from North Carolina. Imran and Peter will try to travel together. Mike will send out a summary email to the pax list.

RE: The Saturday meeting. It would be a work session that would take the ontology that came out of the Friday meeting and work thorough some examples and put them in the ontology. Gary, Mike, and Joanne will be there. If it's just Gary, Joanne and Mike then they may meet before the meeting to work on the examples. We'll play it by ear, Imran will likely be around in the morning and we can meet if necessary.

Aviv can't make the Friday meeting and we don't expect her to make it for the Saturday session. Suzanne will not be attending this meeting.

Agenda for Friday October 3rd Face-to-Face Meeting

Mike sent a tentative agenda that is also available on the project management page of the web. The first item is to spend some time in the morning reviewing the ontology as it stands. Hopefully everyone will be familiar with it, so that this will not take too much time. We will then go through the state subgroup recommendation. After that, we will go through the existing ontology issues that are logged in to the Sourceforge system and any that arise between now and then via feedback from the BioPAX discuss list. We should come up with examples for each issue that illustrate what the issues are. It would be nice if we could have examples prepared ahead of time. We discussed how we could achieve that [JSL after thought... we should have people submit examples with their bug/issue/comment and if not example then the specific bug or screen shot].

We have some existing examples but they're not detailed enough in their current form.

Peter said he could try to translate one of their pathways into BioPAX. The ones that were done before needed to be updated and we need issue specific examples. For example, when we discussed activation and inhibition we needed a specific example that illustrated what we were talking about.

For general examples in BioPAX, it would be reasonable to expect that these could be prepared beforehand, but the issue specific examples Mike mentioned reminded him of what Peter said last time. If we have examples to back up each specific issue, those are harder to prepare because there's so many of them, but if we can, we'll do our best, but some of them we could talk about at the meeting as well.

Peter said if you have a bunch of examples you can usually find a way to discuss problem cases with respect to the examples. Gary said he'd try to put some examples on the Sourceforge issue tracking system. Right now they aren't in there.

The only other issue Chris wanted to talk about was one he suggests to be discussed over lunch or a coffee break. It is how we disseminate, not just release the ontology and examples, but how we actively engage people in putting data into it in this form. Mike will add this to the October 3rd agenda.

Joanne asked if there was anything we could/should do in preparation for the meeting and Chris said one thing we could do is revisit the databases Mike collected. We could go through that and to see which ones of those are most suitable from the point of view of the content and the nature of the database. The preparation would be to come up with the short list of groups that are most likely to actively engage in the process of providing data in this form. And with that (short) list, we could have a concrete discussion. Mike mentioned that we need to discuss what we will do to help users adopt BioPAX.