Summary of the GO Consortium Meeting help March 4 & 5 at the Carnegie Institution, Stanford University, Palo Alto, CA.

Hosts: Sue Rhee and the TAIR group

Participants:

TAIR, SGD, BDGP and FlyBase, MGI, DictyBase, Worm
full list of individuals at the end of the document

Guests:

Han Xie of Compugen
Mark Wilkinson, of the NRC Canada, a TAIR collaborator

Summary Agenda...

This is not in the order of the meeting, but rather supports some structure for this report.

1. **Introduction to GO for New People and Systems.**
2. **Revision of Enzymes to incorporate E.C. terms.**
3. **Short Notes, Updates and Action Items.**
   - **Definitions**
   - **SP2GO**
   - **InterPro**
   - **GO-SLIM**
   - **Obsolete terms**
   - **Energy Derivation**
   - **Top Level Terms**
4. **Sort Process Ontology into component parts and other Process considerations.**
5. **"Determination", "Differentiation" and "Development".**
6. **Major Divisions of the Process Ontology**
7. **Physiology - Initial Discussions**
8. **Report on Narrative vs. Combinatorial approach re anatomy in biological process terms.**
9. **Software Update from BDGP group.**
10. **New procedures for revising ontologies.**
11. **In General, Things to Do, some sooner, some later.**
12. **Specifically, For the Next Meeting (July 14, 15) in Bar Harbor**
13. **Progress Reports inclu. short reports from Compugen and NRC Canada**
14. **Full List of Participants.**

****

1. Introduction to GO for New People and Systems

Brief History, What ontologies are being developed, What are the rules and procedures for both\ontology development and annotation of genomes, Review of the public presence of the GO consortium.

2. Revision of Enzymes to incorporate E.C. terms. It was agreed that incorporating EC higher level terms was a good thing to do. Some of the EC strings are very long because they are adding definitions into the term. We will move the definitions into the definitions file. This will not restrict searches since the search includes the definitions. **Michael will work to tidy the list and replace current enzyme set with new representation.** Synonyms will be added as needed. Curators are reminded that synonyms for the protein should NOT be entered, just synonyms for the molecular function.

3. Short Notes, Updates and Action Items.

a) **Definitions:** We still only have about 10% of terms with definitions. The rule is, if you add a new term, you need to add a definition. We reminder ourselves that the GO:ID goes with the definition, not the term, in cases of revision in the use of a term. SGD crew are scanning OUP Dict of Molec. Biol. into an ascii file for us to use in adding the definitions. This will be incorporated into the GO-EDITOR (John Richter). We will add ISBN numbers to each definition, as well as personal signature to each definition we add.

b) Update of **SP2GO** files. Need to continue with timely updates. New process from MGI will update SP with each MGI update. David Hill (dph@informatics.jax.org) continues to be the primary person managing this file.

c) **InterPro:** Michael Ashburner reviewed history of InterPro for new people. Mapping of InterPro to GO is public at EBI, but is not posted at the GO site yet. **Place InterPro:GO mapping at GO site.**

d) **GO-SLIM**: At the moment, there is a hand-curated GO-SLIM. Ultimately, an attribute of a GO term will be that it is a member of a certain GO-SLIM representation. We recognize that there will be different slices of the GO that will be useful to different annotation communities. So we expect to support different GO-SLIM sets. **GO-SLIM implementation will wait for database.**

e) **Obsolete terms**: When a term becomes obsolete, the definition should be appended to explain why it became obsolete. The note might also contain suggested terms to search if you are considering this obsolete term. John Richter will make sure that obsolete terms are supported in the GO-EDITOR.

f) **Energy Derivation:** Natasha Maltsev of Argonne Natn Lab has list of energy derivations that will be a starting point for expansion in this area. **Michael will get list from Natasha.**

g) **Top Level Terms:** We don't want to limit top level terms. We need to think of them as 'collectors'. So when considering the addition of high level terms, consider 'Do we need this collective term?". When we consider that we have a term (growth and maintenance, for example) because we cannot distinguish by experimental data to which term we should annotate a protein, that is an annotation perspective. But we also want to include terms so that we can group things.

h) **Prions** will not be represented since they relate to disease state.

4. Sort Process Ontology into component parts and other Process considerations.

We discussed whether the process ontology should be separated into two parts: cellular and multicellular. This discussion is not new. We recognized the utility of having a complete unit of the process ontology representing cellular-level processes since this is needed and practical for the unicellular organisms. Thus we will work towards a robust representation of cellular processes that will be useful to all. *This decision led further to a recognition the process ontology is sorting into 4 major components. These are: cellular processes, developmental processes, physiological processes, and behavioral processes.* **We agreed to break out cellular processes and to specifically represent them at the top of the Process ontology.** Most of the discussion over the rest of the meeting then focused on developmental processes.

a.) Differentiation is a cellular process; morphogenesis is a multicellular process

b.) We discussed whether to break apart the terms 'growth and maintenance' and 'cell organization and biogenesis'. At first, it seemed that we should. However, we quickly realized the utility of these terms in that some preliminary experimental evidence couldn't distinguish as to whether a gene product was involved in 'growth' or in the 'maintenance' of an organism. We did agree to **change the term 'cell organization and biogenesis' to 'cell organization and/or biogenesis'.** Midori will incorporate this into the work to carefully define the high level terms. Still some confusion as to the difference between the 'cell organization and/or biogenesis' node and the 'growth and maintenance' node.

c.) **Every high-level node needs careful definitions:** Midori and Michael will work on this soon.

d.) **Remove terms 'oncogenesis' and 'tumor suppressor'.** These terms reflect phenotypes. 'Oncogenesis' is really 'unregulated or mis-regulated cell cycle control.'. The ontology term relates to cell cycle regulation. The evidence for the association of a gene product with the process of cell cycle regulation often come from the study of the disease state. This same argument supports the removal of the term 'tumor suppressor', which is, after all, a phenotype statement and not a biological process statement.

e.) **Cell Motility:** Under 'cell motility', 'vesicle transport' and 'spindle function' are examples of cell motility. So, maybe need to extend upper level with a term 'motility', then a daughter term 'cell motility' and a daughter term of that of 'cytokinesis'. So 'cytokinesis' would be a part of 'cell motility'. **Cell division** is a synonym rather than a GO term because the term is used both as 'division of the nucleus' and as a synonym for 'cytokinesis' , i.e., division of the entire cell. So, **Synonyms need not be unique**. We need some further work here under 'cell cycle' as there are multiple usages of these terms. So, need precise definitions of our usages.

5. "Determination", "Differentiation" and "Development".

Definition of 'Determination' and definition of 'Differentiation'. How shall we represent these concepts? Reflects a 50 year debate in developmental biology. **Need to rewrite these definitions so that they are less experimentally based.** Consider, throughout, 'has the definition been written in terms of the experimental method?', If so, consider revising definition. Sound bites from this interesting discussion

-determination when the decision has been made to adopt a developmental stage (tricky because it is often before the actual differentiation occurs)

-differentiation when you actually express a set of characteristics...process whereby relatively unspecialized cells acquire
-so is '**cell specification**' a synonym for determination? Or is *it that specification is the same as establishing an identity but not yet determined*. It is a temporal thing. you are getting signal. it is not the same as determination.
-autonomous specification specification produced by a inheritance of molecules. a type of cell specification
-conditional specification is the specification determined by the relative position of cells in an organisms. A type of cell specification.
-not the same as **competence** which is a characteristic of a cell.

Conclusion: Competence is the 'ability' to do something. Competence is not a process, it's a state. So we throw it out. but, if useful, we could have 'establishment of competence' or 'maintenance of competence'

Conclusion: the term 'Development' as a high level process will be used to consider the whole history of the organism.

This generated a lot of discussion as we considered 'embryonic development' and 'post-embryonic development'. This is a hard distinction to support for plants and for larval development. Different communities use these terms in different ways. Post-embryonic development is useful for fly...keep it in???

What is covered by the term 'brain development'? It continues throughout life. What do we mean by a term like 'heart development'? Does that mean the developmental process up until you have a heart? Or does it include the further development of the heart after a recognizable organ is formed?

Embryogenesis, morphogenesis, organogenesis are all DAGs...some things that parts of embryogenesis will be part of morphogenesis as well. So...

development
- morphogenesis
  - aggregation
- differentiation
  - maturation
- aging
- senescence

Option 1 global heart development
- formation of the heart
- beyond formation of the heart

6. Major Divisions of the Process Ontology.

% cell
% development
% physiology
% behavior

7. Physiology - Initial Discussions. Having struggled through the beginnings of a representation of development, and at least conceptualizing the work needed to realize this part of the process ontology, we recognize that physiology is the next big area to struggle with. Animal and plant physiologies will be pretty independent. Cellular processes are independent of physiology. So here is where the DAG structure becomes imperative. For example, 'hormone response' has both physiological and cellular components. We can relate them through the use of the DAG structure.

Remember that we are trying to develop a tool for biologists that works....not trying to represent all biology. Need to make somewhat arbitrary decisions, such as where to put 'germination', that address what the user cares about, i.e. 'what genes are involved in the process of germination?'.

Physiological processes are heavily impacted by outside signal (environment). Changes in response to environment. While not absolute, physiological statements more often reflect processes in the mature organisms.

In defining physiology as a grouping mechanism, we need to work down to the next level now. 'Transpiration' is an example of a physiological process, ditto 'perception of external stimuli', 'stress response', 'immune response', etc. 'Seed germination', 'release of dormancy' terms are both physiological and developmental processes. Physiological processes ultimately will go down to the granularity of cellular processes. The DAG structure will help in the representation of all these terms in the Process ontology.

8. Report on Narrative vs. Combinatorial approach re anatomy in biological process terms.

This was the major event of this meeting. For many meetings, we have come back to the issue of species-specific anatomies and the incorporation of anatomical terms in the process ontology. Over a year ago, Joel Richardson proposed a combinatorial approach wherein a process term combined with an anatomical term would be used to annotate knowledge about a gene product. At the Hinxton meeting, the group agreed that this was a sensible and powerful approach. However, subsequent implementation efforts revealed difficulties in incorporating such biologically useful concepts as 'gametogenesis'. Also, the management of the combinatorial approach would be harder than the further development of what is now call the 'narrative' approach. The narrative approach is the current paradigm of building up the ontology incrementally as we describe the process in biological terms. Yet, in following discussions, the issue of whether or not to incorporate anatomies, which are themselves highly developed and precise ontologies, in the process ontology kept arising. Finally, at the human annotation meeting at Banbury last summer, we agreed that David Hill would 'do the experiment' and give a presentation at this meeting for the group to consider.

David used the example of "Heart Development". He developed ontology for heart development in both the narrative and the combinatorial manners. A copy of that presentation is available. The end result was that the group was overwhelmed with the power of the combinatorial approach both to provide self-structured cross-product terms and to reveal new information and avenues for experimentation.

1. Do we leave it up to each group to decide whether to use this approach to process annotation? A resounding NO from the group.
2. Can we separate out subtrees that can be used to generate cross-products? Yes, could use GO-SLIM or other subtrees. In fact, the GO-SLIM set may be the mechanism for grouping annotations across species.
3. There could be cross-products of cross-products....how far do we want to break this down? Don't have to go all the way down as long as the representation of the biology is correct.
4. Works as long as the two concepts are orthogonal, can't do with just anything and get the consistency needed.
5. Big worry...if each group is incorporating combined terms relative to their particular anatomy, we lose the power of the combination of all annotations. One approach is to ask the query...'give all products in heart development', and have query go out against all cross-products. We will have to work on this.
6. Can we have a join of the anatomies? then have a single anatomy to use in the cross-product with developmental processes? don't know...right now, we think the combinatorial approach is the right way to go, we will have to work on the implementation.
7. Some concern about ripping out anatomical terms from process right now. Can the primary process ontology be made more amenable to cross-species specific anatomical parts?
8. If we have multiple anatomies, then the search needs to go against anatomies...this can be done.

Summary....Issues
1. There is general consensus to go forward with the combinatorial approach.
2. Do we need to have a shared anatomy?
3. How will others be able to use the ontologies to annotate if we have this complicated approach?
4. Parser...need to put into better language...earlier we tackle the problem of language, the better we can promote this for ourselves and others.
5. GOAL...write definitions for common developmental process terms.
6. Start working on further experiments with this approach...write definitions, work out mathematical properties.
7. Each group needs to provide an anatomy.
8. The anatomies needn't have GO:IDs, but the cross-products should have IDs.
9. We will use the developmental process as a demonstration of this approach..
10. Immediate action items include:
    a. schema changes (Joel and Suzi)
    b. editor will work fine for now.

9. Software Update from BDGP group.

Brad Marshall - GO-Browser. Objective is to make browser better. Have moved from cgi scripts to XML backend with RDI to associate with different data sources. This makes for a more flexible backend. Want to chain a lot of data sources

together with GO associations. Only want to retrieve a subgraph at a given depth. Much enthusiasm for power of this approach.

John Richter - GO-EDIT. This new editor is an open-source application that provides an annotation tool for GO type ontologies. Can rearrange, define terms, designated subsets... Released and available. Will start using this right away. First commit will result in messy Diff file, but then all will be well.

1st change....editors using Editor will write out to files and commit via CVS

2nd change...editors using Editor will write out directly to database...don't know when that will be.

John Richter - GO database. database could be ready 3 wks after he starts working on it, but right now he's working on Apollo.

Suzi Lewis - Apollo Pedigree

10. New procedures for revising ontologies.

* who has 'write' access in each instance?

Michael, Heather, David, Harold, Leonore, Midori, Karen

* how are people outside suppose to communicate suggestions to us?

Suggestion...two databases, one public, one writable (production). Curators will have db accounts...login to edit and write. A few have publish access. Publishing takes it over to the public db. We track changes, etc.

Publishers.... Michael, Heather, David, Harold, Leonore, Midori, Karen(Publishing involves clearing and reloading. The event gets a version number.

'Static' version would be a cron job (midnight of the first of every month???) There would need to be a name for each release.

11. In General Things to Do, some sooner, some later

1) Cross Products: Post GO:IDs for cross-products not for Anatomy...

2) Details for Cross-Products...may be summer before we can commit all this.

3) Post InterPro:GO mappings at GO Web site (Michael)

4) Post Anatomy Files from different groups.

5) Post FASTA files of unique set of AA seqs with GO annotations at GO site, include SP:ID in header. Set up for searching.

6) Review and update GO-SLIM files

7) Develop top levels of Physiology for next meeting

8) Add seqID column to Gene Associations table (use SP ID).

9) Consider posting 'good' other ontologies at GO site. This will involve a lot of discussion...Don't want impression that GO is responsible for quality of other ontologies.

10) Update SP translation tables (David Hill)

11) Move comments about reasons for changes from CVS to database (when we can) (John Richter).

12) xml dumps from Editor to Suzi.

13) Support old GO:IDs in database (John Richter).

14) Post 'Citing these data' on Web pages (MikeC).

15) Provide some kind of static or versioned GO for use by tools that incorporate the GO as part of their annotation suite.

16) Suzi needs to deal with <>...where ever you want to keep this character, put the / in front of it.

17) MikeC will create an anonymous server behind firewall at Stanford for cvs or provide a machine outside the firewall.

18) Make CVS world readable..provide a repository..may want to consider SourceForge...

12. Specifically, For the Next Meeting (July 14-15, Bar Harbor).

1. Midori and Michael will have some High-Level definitions for review (i.e., when does the process of differentiation start? when does it stop?). **Change the term 'cell organization and biogenesis' to 'cell organization and/or biogenesis'.** Also, need children of combined terms....thus, need 'cell organization' term and 'biogenesis' terms with definitions.

2. Need some work done to clarify and define terms in the area of 'Cell cycle'.

Full List of Participants.

TAIR: Sue Rhee, Leonore Reiser, Aisling Doyle, J. Yoon, Margarita Garcia

DictyBase: Rex Chisholm, Warren Kibbe

SGD: Mike Cherry, David Botstein, Midori Harris, Selina Dwight, Karen Christie, Dianna Fisk, Anand Sethuraman, Cathy Ball, Gavin Sherlock,

Worm: Wen Chen

BDGP and FlyBase: Michael Ashburner, Suzi Lewis, Heather Butler, John Richter, Brad Marshall, Chris Mungall

MGI: Judith Blake, Joel Richardson, David Hill, Martin Ringwald, Janan Eppig, Harold Drabkin