GO Consortium Meeting – Stanford, CA – January 15-17, 2004

[Next Meeting: Chicago – Dictybase organizing – October 2004]

Group Participant List

SGD	(Mike Cherry, Karen Christie, Kara Dolinski, Eurie Hong, Dianna Fisk, Rama Balakrishnan, Rob Nash, Stacia Engel)
TAIR	(Sue Rhee, Tanya Berardini, Suparna Mundodi)
MGI	(Judy Blake, Joel Richardson, Harold Drabkin, David Hill, Mary Dolan)
ZFIN	(Doug Howe)
RGD	(Victoria Petri)
Dictybase	(Rex Chisholm, Petra Fey, Karen Pilcher)
EBI-Ontology Group:	(Midori Harris, Jane Lomax, Jen Clark, Amelia Ireland)
GOA	(Evelyn Camon, Daniel Barrell)
Wormbase	(Kimberly Van Auken, Ranjana Kishore)
Incyte	(Burk Braun)
Gramene	(not present)
IRIS	(Richard Bruskiewich)
Berkeley DB Group:	(Suzi Lewis, Chris Mungall, John Day-Richter, Brad Marshall)
TIGR	(Linda Hannick)
FlyBase	(Michael Ashburner, Rebecca Foulger)
S. pombe/Sanger	(Val Wood)
Pathogen/Sanger	(not present)

TOC

1. Opening Comments: GO Grant Update (Judy)

Judy reported on the status of the GO funding from NHGRI. In the competitive renewal, the GO funding mechanism was changed from an RO1 to a P41 (research_resource), and significant new funds were requested. Current indications are that we will be funded for 3 years. However, there have been several cuts in funding proposed including one requested software engineering position. Additionally, there will be no new group (sub-contract) funding. Small side projects are also not funded. There may be additional adjustments; we are awaiting official notifications. We hope that any further adjustments in funding can be shared across all of the groups receiving funding through this grant except for the European contracts (due to dismal exchange rate) and BDGP (which has already had one position cut).

2. Annotation Progress Reports:

Reports were issued from the following groups:

SGD   -   Report available
TAIR   -   Report available
MGI   -   Report available
TIGR   -   Report available
ZFIN   -   Report available
RGD   -   Report available
Dictybase -   Report available
Flybase -   Report available
GOA   -   Report available
EBI-Ontology -   Report available
Wormbase -   Report available
Incyte   -   Report available
Sanger/Pombe -   Report available
Sanger-pathogen - Report available
Gramene -   Report available
IRIS - ok, no electronic report
BDGP Software - Report available

3. Interest Group Reports

a. Plant Interest Group
Plant interest group report is available at
http://www.ebi.ac.uk/~jclark/GOwebsite/text%20in%20development/plants_folder/plants.htm

This is also available as a text file with other reports from this meeting.

No other interest groups reporting.

4. Ontology Development Issues
4.1 Metabolism terms

It was decided that "metabolism" would be split into "cellular metabolism" and "organismal metabolism". This is similar to the division of "physiological process" into "organismal physiological process" and "cellular physiological process". Further discussion about this will continue in coming weeks.

4.2 Regulation of non-biological processes

Example: regulation of water crystallization: water crystallization is not a biological process but it is regulated biologically (*e.g.*, "regulation of water crystallization" IS a biological process). Some don't have "is a" parents. How

about "water metabolism", homeostasis**?** **Action item**: find homes for these terms. As we proceed with consistency checks throughout the ontology, we will need to provide parents for these terms.

4.3 Transcription/translation factor activity

Many of these terms are really describing complex processes and not activities. For example, "translation initiation factor activity" (GO:0003743): functions in the initiation of ribosome-mediated translation of mRNA; "transcription factor activity" (GO:0003700): Any protein required to initiate or regulate transcription; includes both gene regulatory proteins as well as the general transcription factors. There is not a specific "activity". These terms **are** heavily used by the biology community. What is needed are real definitions. Where should a regulatory activity go? One suggestion is to think about how a particular activity is being assayed to help with making a real definition.

There was a long discussion about this. From a practical point of view, not all activities are usefully 'atomized'. Biologists may not be able to nimbly provide a definition of 'transcription factor activity', but they do have a conception of the complex matrix of action described by this term. We agreed that these more complicated function terms need to be included in the ontology.

>   **Action item:** Eurie and Michael will strive to provide a definition for 'transcription factor activity'.

4.4 Component ontology annotations
>   Transient associations for complex: "associated with" vs "part of". Should we distinguish between stable components of a complex versus something that by some experiment localizes to the complex?

>   **Action item:** We will add a new qualifier for "Colocalizes with" that is appropriate for indicating that the gene product has been found in the vicinity of a structure.

>   **Action item:** The GO office will update the documentation for Component rules with discussion of this qualifier and its use.

4.5 Protein classification:

The use of a protein classification system in the GO is being investigated.

4.6 Sensu definitions:

Jen lead the discussion of various options for incorporating taxonomic information as necessary

>   **Action item:** 'sensu' terms will have a mixture of English phrases and Latin genera, along with the taxon ID. The definition of any sensu term would include the point that it is not totally restricted to a particular grouping.

http://www.geneontology.org/GO.usage.html#taxon

4.7 Documentation of function ontology

Current documentation for creating terms for the function ontology is being further refined. One item that was discussed was representing complex functions in function (e.g., receptor tyrosine kinase, which is a receptor (ligand binding) and a kinase) (doesn't make sense to deconstruct); if "receptor tyrosine kinase" is a child of both receptor and kinase, then its function is defined by its placement in the graph. In some cases, however, there may be a complex function: when the two or more functions are not mutually exclusive; the two activities are **coupled;** the functions are dependent on each other.
>   **Action item**. These concepts will be added to the documentation.

4.8 GO_Slims Development

Some people wanted documentation for developing their own GO slims. Currently, there is only the readme file in the GO-Slim directory in cvs. There IS software:

Map2Slim script: Takes an annotation and puts it into an appropriate place in the Slim. To use, you need a Slim file and association file. The output result is a binned output. The program allows creating bins that don't exist in the GO; like chaperone + Chaperone regulator. This is available at GO site.

4.9 NameSpace Ontology

Mark Wilkinson will be responsible for updating BioMoby and with provided NameSpace designations.

4.10 Improve GO email archive search?

A suggestion was made to have the GO email archive searchable by Google. However, it would mean that everyone in the world will see it. If we can restrict it, that would be better.
> **Action item:** Brad and Mike will look into whether it is possible to keep a Google search of the email archive separate from the general Google search of the GO web pages.

4.11 Gene Association file errors

Each group has checked these. However, one issue brought to light is that the files are getting big! It was suggested that we start compressing these files (gzip). Also, many are not updated frequently; in May two more will be over a year old. The groups need to address this.

Also, it was suggested that people having write access to the cvs repository use ssh (so that password is never free text over internet). The downside is that one has to type password more often; however, it is more secure. John pointed out that we don't need to type our passwords each time if we get a private key on our computers.

Action item: groups to investigate if large files, compressing files, will pose any problems at their own site.

4.12 Tracking the date of a term definition

A suggestion was made to track when a term definition was made. The date might be added to the definition itself, perhaps through DAG-Edit. **Action item**: Implement if easy and does no harm.

5. Software Report
5a. Chris Mungall gave a presentation of OBOL (**Open Bio-Ontology Language)**; this is available at the GO site (http://www.geneontology.org/meeting/go_2004_01_stanford/). This is the result of the decomposition of implicit info out of GO terms into defined classifications. OBOL will support the slot-based annotations discussed in previous meetings, and will more readily support creation of cross-product and composite terms.

5b. John Day-Richter summarized the DAG-edit workshop held at Hinxton in Dec. 2003. It was very useful for working out bugs and highlighting functionality that many people were unaware of, such as
   a.   CVS plugin for DAG-Edit
   b.   Term change tracker plugin
   c.   OBO data adapter: can break ontology into multiple files based on name spaces
   d.   Category manager plugin: used for GO_Slim set up

Future enhancements will include include filter plugin, search, color, decoration, and more.
Some of these, however, will require switching to OBO format files.

Action item: Proposal: by end of month will post OBO files, curator trials -- use OBO for two weeks to work out bugs, then general switch to OBO as master; More general document for users. Announcement on site

Why use new format? The advantage of new format for average user is that the file is more readable; it is one file rather than four; it is smaller overall (35%), and using the OBO format will allow the user to take advantage of many of the functions of DAG-Edit that are not saved with the old flat file format.

5c. Brad Marshall reported that although the AmiGO browser seems to be the same, many changes have been implemented to the way it functions. There are also plans to incorporate GraphViz visualization. The GO database is currently being updated monthly. However, once it is moved to Stanford, the updates will be more frequent.

6. Annotation Issues

6a. Problems with pathway information annotation; inference from existing genome annotations to novel genome annotations.

A discussion was generated by TIGR (Linda and Michelle Gwinn) concerning how to annotate pathways. GO does NOT, per se. The following example was discussed:
A is converted to B via x (gene product); B is converted to C via y (gene product)
X in annotated using the GO ID for its function; Y is annotated to the GO ID for its function (both need an ISS at least); Then, if there is a Process term Biosynthesis of C, both gene products are annotated to its GO ID using an IC code (the curator infers the pathway because both x and y exist. The with field when the IC code is used here will have at least two GO IDs corresponding to the two GO activities (read: the curator infers that this process occurs because of the presence of the two activities (GO IDS) in that organism.

7. Future Meetings

It has been brought up several times that due to the increasing size of the consortium as well as budget considerations, the usefulness of the large group meeting every three to four months is less than it was. It is suggested that these large group meeting occur no more than twice a year. However, there will be planed smaller meetings with a specific focus. The intended participation would be limited to those members that have a specific interest in the topic.

Suggested topics would be:
- Tools and development (*e.g.,* such as the DAG-Edit workshop)
- GO Annotation **(This might be open to other groups like Xenopus, etc. This might be run as annotation jamborees, etc. Might be a good place to discuss quality and efficiency problems, etc.)
- SO (sequence ontology)
- Ontology Content (*e.g.* Tanya and David for cell physiology)

When a large meeting occurs, it will be important that someone that goes to the smaller meetings attends to give a report of these meetings to the group as a whole.

8. Final Item.

Worm Book III would like to embed use of GO terms within all articles of the book; the online version of the book would then link out the appropriate GO term pages . This is a good idea for us to be involved in, especially with respect to our grant mandate to make GO more accessible. Perhaps we could make a dictionary based on an alpha dump of the ontology. We could post it periodically. Obsolete terms would need to be removed.
ACTION ITEM: Mike Cherry to look into how best to do this.

9. Summary of Action Items from this meeting.
1. Action item: Eurie and Michael will strive to provide a definition for 'transcription factor activity'.

2. Action item: We will try to set up a pilot project that has a web page "indexing" key point discussions in the GO email archives

3. Action item: We will add a new qualifier for "Colocalizes with" that is appropriate for indicating that the gene product has been found in the vicinity of a structure.

4. Action item: Jen will update the documentation for Component rules with discussion of this qualifier and its use.

5. Action item: Brad and Mike will look into whether it is possible to keep a Google search of the email archive separate from the general Google search of the GO web pages.

6. Action item: groups to investigate if large files, compressing files, will pose any problems at their own site.

7. Action item: Mike Cherry to look into how best to interact with WormBookIII to embed GO terms in on-line version of the book. Need to consider 'glossary' approach and how to maintain currency.

8. Action item: 'sensu' terms will have a mixture of English phrase and Latin genera, along with the taxon ID. The definition of any sensu term would include the point that it is not totally restricted to a particular grouping.

9. Action item: Proposal: by end of month will post OBO files, curator trials-- use OBO for two weeks to work out bugs, then general switch to OBO as master; More general document for users. Announcement on site

10. Action item: We will try to set up a pilot project that has a web page "indexing" key point discussions in the GO email archives.

11. Action Item: "Not" column will be renamed "Qualifier". When it has any other value other than NOT or NULL, it should be used for annotations for components of a complex only. This will allow reason across membership in a component to infer function. Should be checks for complex entries. Subunit will have annotation to a particular subunit activity, if known, or to "contributes_to" and that gene product must also be annotated as a component of complex.

e.g., specific example eIF2; has three subunits (alpha, beta, gamma); one binds GTP; one binds RNA. But the whole complex binds the ribosome (needs all three); so all three get "contributes to" ribosome binding, and one gets GTP binding, the other gets RNA binding. AND all three are annotated to EIF2 complex.

12. Action Item: In column #12 of the gene_association file, "complex" will be allowed as a type of "DB_OBJECT".

13: **Action Item:** Concepts relating to the use of complex functions (e.g. receptor tyrosine kinase) will be added to the documentation.

14. **Action Item**: Add Joel Richardson's tool to the tools page. Note language change from Python to Java. (Jen)

15. **Action Item**: Document OBO flat file format advantages for annotators (There are none.)

16. **Action Item:** Write documentation for the process and component ontologies along the same lines as the function documentation that had already been written. (Jen)

17 Action Item: Add documentation to remind people that the definition is there to clarify the meaning of the term name if there is any ambiguity. This is to be added to the general documentation as well as to the documentation for each ontology. (Jen)

18. Action Item: See if there is an easy way to add the date that a definition was made.

10. Review of Action Items from past meeting [Bar Harbor]

1. Standard operating procedures for checking ontology integrity: These would include checking true path rules, missing parents (terms that are not "is_a" to anything), etc. *Currently, some groups have scripts that do this; perhaps these can be incorporated into DAG-Edit. This will be continued.*

2. Documenting process for revision of sub-trees (for example, the changes being made to split physiological process. *Currently much of the discussion occurred via telephone. These type of thing should be documented so new members can review discussions – We need some sort of "audit trail" of discussions – some are in SourceForge entries , and should be made easier to find. It is important that more than just changes but also the rationale and logic about things that have been extensively discussed. Ultimately, summaries should be submitted to Sourceforge. A suggestion was made that chat room technology could be used instead of phone conversations, so that everything is logged.*

   Action item: we will try to set up a pilot project that has a web page "indexing" key point discussions in the GO email archives.
3. Work on documentation on procedures for getting people into interest groups *is progressing*.
4. The fate of "signal transducer activity" is still under investigation.
5. The checking of logic consistencies for "Part-Of" is *in progress*
6. When is a function actually a process? *Documentation is being drafted to add to principles of ontology development.*
7. A Source Forge item on "regulation of survival gene products" under "apoptosis" to be checked by GO team. The term name was changed to "regulation of surival gene product activity" and moved under "anti-apoptosis". **DONE**
8. Synonym types: *Documentation describing the types of synonyms (exact, related, broader, narrower, not same) is on the GO site. However, synonym type is only displayed in OBO format files.*
9. Added English terms as synonyms as needed. **DONE**
10. Resolution of the "regulator subunit of enzyme activity" as a child of the enzyme activity: The enzyme activity will no longer have its regulatory subunits as children. The term "enzyme regulator" will be the parent of these terms. **DONE**
11. Additional use of the NOT field .
The discussion basically revolved around three choices:
   a. Allow NOT field to have multiple use. It was suggested that we rename qualifier column from "NOT" to "qualifier", and can have values "NOT" or "contributes_to" or NULL. If using "Contributes_to" to annotate to a function term, then the annotation must be accompanied by a component annotation for same gene product to the complex. There was some concern about putting two different types of info in one column.. Subunits are annotated with function of complex as "contributes_to" as qualifier and subunit annotated as component of complex.
   b. A special evidence code for being in a complex instead of "contributes to". Like CPX?
   c. Use an additional column to indicate that the activity occurs only when complexed to something else.

Action Item: "Not" column will be renamed "Qualifier". When it has any other value other than NOT or NULL, it should be used for annotations to component only. This will allow reason across membership in a component to infer function. Should be checks for complex entries. Subunit will have annotation to a particular subunit activity, if known, or to "contributes_to" and that gene product must be annotated as a component of complex.
   e.g., specific example eIF2; has three subunits (alpha, beta, gamma); one binds GTP; one binds RNA. But the whole complex binds the ribosome (needs all three); so all three get "contributes to" ribosome binding", and one gets GTP binding, the other gets RNA bindings. AND all three are annotated to EIF2 complex
12. Two new curators added to web site "people" page: **DONE**
13. New web site with improved index implemented. **DONE**
14. Changes to function ontology documentation: **DONE**
15. Document difference between parent/grouping in function vs a single term in process. **DONE**
16. Different types of "Part of". OBO can handle, but flat file format can't. *The are 5 different types of ' part of' are documented.*
17. Document that "binding" is not always a parent of enzyme. It is only a parent when stable binding occurs. **DONE**
18. Standardization of the use of "activation", "induction", and "positive regulation of" *is being documented.*
19. Add any procedural information coming out of Curators/Annotators meetings
20. Bad link was fixed in GO.evidence.html **DONE**
21. New Tools: Joel's Graphical Annotation Browser demo.

22. DAG-EDIT 1.409 beta 4 being tested.
23. Synonym documentation completed. **DONE**
24. Document the use of the NOT field for an ISS annotation with sequence dissimilarity. **DONE**
25. Document that the with field used with ISS can have cardinality > 1. **???**

26. An "Annotation Oversight Team" will be created to assess quality, set standards, evaluate annotations of contributing groups, and alert groups to annotations that need attention. *Not Done: Will continue with plan to have 'GO Annotation Camp' this year.*
27. COMPUGEN will resubmit their annotation file. RESOLUTION: If there are IEA sets that have not been updated in one year, they will be removed from the front page and AMIGO.
28. See 24.
29. **Action Item**: In column #12 of the gene_association file, "complex" will be allowed as a type of "DB_OBJECT".
30. A tool is needed to check validity of an annotation that was made to a Riken gene based on ISS to a SP record or IP domain. When the IP domain is removed by SwissProt (because it was actually found to not be a real domain, or always associated with a particular activity, etc.) the annotations are now no longer valid. *This is being researched by Suzi and David.*
31. Documentation added to tell annotators that if you are unsure of function/process, bump annotation up to next level; it that results in the root of the ontology, you need to use "unknown". **DONE.**
32. Formatting errors in association files are being found: everyone should check. **DONE**. *A proposal was suggested that all use the same script to check this.*
33. Getting groups references for any "in-house" analysis and post on GO site *continues.*
34. Annotation QC: defer to annotation discussion
35. Send out reminders to groups to update their gp2protein files! Used in Amigo and other.
36. A production manager that will work on AMIGO, GO database, etc. will be hired
37. This is #27 again.
38. AMIGO will have a filter added for "organization" and "species"
39. Sourceforge now has a tracker for AMIGO requests.
40. Format for IEA references *in progress.*
41. Term change tracker plugin in DAG Edit allows you to track changes. DONE
42. A second DAG-EDITOR course will be held on the West coast..
43. New flat files (OBO format) will have "obo" extension. At some point, the three ontology files will be combined.
44. Action Item: hire a person to run test sets on all of the GO tools.
45. Request to translate the GO into various non-English languages, including Arabic and Chinese. Pankaj to investigate the need, how to update, etc.
46. Create documentation for ontology development guidelines aimed at people that we would recruit as outside experts to develop specific branches of GO. EBI: in progress
47. Mike Cherry developing manual methods, automated assessment tools and documentation aimed at improving annotation consistency.
48. Improve consistency in annotation. Mary Dolan is currently analyzing the annotation consistency between mouse and human genes. ; continue working;
49. The "decomposing" of the GO ontologies is in progress. To be reported on at a later date.