**GO Meeting July 14-15,2001**
Hosted by Judy Blake and the Jackson Laboratory in Bar Harbor, ME.
Minutes compiled by Leonore Reiser

**Attendees (name-organization):**
Judith Blake- MGI
Janan Eppig-MGI
Joel Richardson-MGI
Martin Ringwald- MGI
David Hill- MGI
Harold Drabkin-MGI
Michael Ashburner-FlyBase
Midori Harris-GO
Suzi Lewis- BDGP
John Richter- BDGP
Brad Marshall-BDGP
Pavel Tomancak- BDGP
Mike Cherry- SGD
Anand Sethuraman-SGD
Karen Christie-SGD
Selina Dwight- SGD
Dianna Fisk-SGD
Erich Schwarz-WormBase
Raymond Lee- WormBase
Rex Chisholm- DictyBase
Liat Mintz- Compugen
Courtland Yockey-Astra Zeneca
Rolf Apweiler- SWISS-PROT
Nicola Mulder- SWISS-PROT
Jennifer Hogan- Incyte
Matt Berriman- Sanger
Tom Weigers-MGI
Jim Kadin- MGI
Carol Bult- MGI
Sue Rhee- TAIR
Leonore Reiser- TAIR

**Progress reports by group (presenter):**
**MGD (Harold Drabkin):** from handout supplied.

1. As of 7/12/01 MGI has 15638 annotations.
2. Total number of mouse genes with at least 1 associated GO term: 5673
   Genes with molecular function term: 4615
   Genes with process term: 3487
   Genes with component term: 3584
3. Breakdown by annotation type:

IEA: 3690 process; 5333 function; 3835 component
Of these: 760 annotations are from EC mapping (694 function; 66 component) and 6542 from SWISS-PROT mapping (1740 process, 2603 molecular function; 2199 component). The remainder are from GOFISH.

Hand Annotations: Annotating at a rate of 40-45/week. 1115 genes annotated this way (636 process; 637 component; 788 function) that correspond to a total of 2780 annotations (1058 process; 722 component; 950 function).

**SGD (Karen Christie):**

1. How SGD deals with annotations to unknown (function, component or process) is by assigning a reference, from either curation or from a publication with the appropriate evidence code. This distinguishes between cases where a curator has looked and nothing is known and cases where the literature for the gene has not yet been examined in order to make a GO annotation to a given ontology. We decided that this was a good approach for all curators to take and that this decision would be reflected in the documentation on the website.
2.Validation checking: includes checking that all of the required fields were being filled and scripts to identify terms that become obsolete so that annotations to these terms can be updated.
3. Cathy Ball and others have done a four- way genome comparison- Arabidopsis, Yeast, Fly ,Worm:- each sequence is BLASTED against each other to build gene families Criteria is (P =50 and 80% in one HSP).Each sequence is present only in one tree. Compared annotations with a high node, GO Slim type, process ontology and curated each of them by hand (by sequence similarity and FASTA definition lines). Annotating the entire gene cluster (have 1400 families- hand curated for process–14/15 very high level processes, 23 families that could not be ascribed to a process). Function calls were made electronically via a script. This dataset can be used as a check for annotations- and as a general guide for annotation. A limitation is that the represented proteins are only those found in all genomes.

**Flybase Report (Michael Ashburner):**

Working on cleaning up existing annotations:

1. Heather was fixing all legacy annotations that did not have literature or evidence codes.
2. Go Slim high level annotations being gone over for 7000-8000 genes. Michael has gone through all of them- annotated by hand .
3.Several thousand BLAST results being gone through.
4. 9000 genes with at least one GO annotation.

**WormBase (Erich Schwarz):**

1.One third of *C.elegans* genes have been annotated using InterPro.This data is up at the GO site.
2.For 19,000 cds sequences, ~2,200 have ‘reasonable’ phenotypes from RNAi screens. Working on turning phenotype data from RNAi to GO annotations. Mapping phenotypes

to GO. For example the trait, paralyzed phenotype would map to GO term: locomotory behavior. Evidence inferred from RNAi. (which would be given the evidence code IMP- this additional definition of IMP will be reflected in the documentation on the web pages.

**DictyBase (Rex Chisholm):**
1.Initial annotation from genome sequence 5X coverage ~8000 ORFs.
2. Have about 50 hand annotations. Will use InterProt, SWISS-PROT, EC mapping and have a set of annotations by next meeting.

**Compugen (Liat Mintz):**
1. Concentrating on GO Annotations at the level of transcripts.
Methods include:
a. Protein clustering (Smith-Waterman)
b. Literature Clustering-text mining tool.
c. mRNA clustering

First release includes human, rat ,mouse annotations. Update to be released Aug.1 will have additional annotations.

**SWISS-PROT/InterPro (Rolf Apweiler):**
1.Nicola Mulder's map of InterProt to GO terms has about 2700/4000 terms mapped. Many can't be mapped yet ( as they include lots of viral terms).
2.SWISS-PROT mapping. From this about 50% of the terms have been mapped to GO but this still needs some manual curation.
3.Human gene annotation status. Yesterday (July 13,2001) the first pass annotation from SWISS-PROT, trEMBL and Ensembl annotations was completed using three electronic mapping methods. Imported 7316 GO annotations from Proteome and literature associations. The plan is to have complete coverage at a high level by September-October. After this , manual GO annotations will be part of the normal curation pipeline.

**GO (Midori Harris):**
1. Training SWISS-PROT curators to do GO annotations.
2. A bit of PR work with Nature and the Wall St. Journal.
3. Working with Karen Christie of SGD to revamp the GO web pages.

**TAIR (Leonore Reiser):**
1.3901 annotations to genes corresponding to TIGR open reading frames using term matching (exact matches between GO terms and definition lines). Matches generated by script and validated by curator went in as IEA annotations to GO and TAIR FTP site. All annotations are to molecular function.
2. Progress on development of an anatomy ontology for Arabidopsis using GO editor to make DAGs. 266 terms as of 7/12/01, 50% have definitions Will work with rice (Gramene DB, IRRI) and Maize (MMP) to find top level terms for plants.

**Discussion Points:**

**1.Go Slim**

a). Consensus that there needs to be a new GOSLIM developed. Terms for GOSLIM will be selected by a small working group.

b). A directory of the GOSLIM versions that have been used should be made available via the website.

c). Some considerations in using GOEDIT to make GOSLIM files: Will have to wait until the database is up to implement GOSLIM notations as this is not accommodated by the flat files. Also, having everything in the database will make it easier to keep GOSLIM in synch with the current GO. The 'canonical' GOSLIM will be in the database and other versions (specific to certain projects) will be posted as flat files.

d). Chris Mungall has been working on software for mapping full GO to GOSLIM.

e). Midori Harris will take charge of new GOSLIM.

**2.Gene association files**

a). Should we get rid of aspect in the files? Consensus is no, as this information is useful in consistency checking.

b). How to deal with the fact that each group is annotating at different levels/to different database objects. At the moment, most groups are annotating at the level of gene, or transcript, but this differs from database to database. There is a need to define the object being annotated explicitly. Changes to be made are: 1) add a column that defines the object being annotated (or the moment the options will be gene, transcript and protein). 2) The symbol used in the association file will be the symbol for that object (e.g. if annotation is to a gene object then symbol = gene symbol, annotation to protein object then symbol= protein symbol). Same holds for synonyms, they should match the object being annotated, 3) add a column for Taxonomy_ID (from NCBI) that defines the taxonomic node for the organism whose gene/protein/transcript is being annotated. Symbol is still mandatory but since not every database has symbols for everything. In that case, using alternative names is OK (e.g. a gene symbol when annotation is to a transcript or protein. There may need to be further discussion on the issue of symbol column.

c). Should there be another column specifically for PubMed references in addition to the database references for the evidence for the association? No, but we can allow for multiple entries in the DB:reference column to accommodate. Multiple reference identifiers (e.g. |MGI:209393| PMID:123333) will be separated using pipes with the model organism database identifier preceding the secondary identifier.

d). What to do with sequence identifiers? We reaffirmed a previous decision NOT to put GenBank or EMBL identifiers in the association files. Instead, sequence identifiers will be provided in a separate file.

e). Action Items**:** Midori will update the web pages with the new information/format for the association files. An XML format will be created to export the association files.

**3.Definitions status:** Up to 13% of terms are defined. Everyone agrees it is harder to define the higher nodes but crucial especially as these are used for GOSLIM. Midori has taken a stab at some of the higher nodes in process and these need to be looked at. With respect to making it easier to add definitions, SGD has converted a dictionary into an electronic format.

**4. Discussion of "is this a function ?".** Many people have noted that there still seem to be gene products in the GO both in function and process ontologies. Dyenin was cited as an example of a protein name in the function ontology.

a. Rex will look into identifying areas that need to be cleaned up as far as protein names and bring the suggestions back to the group.

**5. How granular should we get in the ontologies- what belongs?** The discussion about what is a function raised questions about how granular should the ontologies get and what is a function vs. a process term.

a). Granularity. In general, the answer is as far as we can go, within the bounds already defined. Community feedback into the ontologies is especially important as a means to improve granularity.

b). What to do with molecules?

If we want to represent the function of a subunit ( a molecule).

1. Add parts of the complex (e.g. add regulatory/catalytic component)
2. Remember we are annotating the product to the potential of the component.
3. To avoid a proliferation of terms, we can collapse nodes and gain specificity by the combined annotation. For example, RNA polymerases might be collapsed into a single function node and the component or process gives the specificity.

c).What about phosphorylation? Is was suggested that having phosphorylation as a process was redundant with kinase in function and is inconsistent with the rule that a process requires more than one function. In this case we decided to keep protein phosphorylation in process ontology because it would be expected as a child of protein modification.

**6. Status of the GOEditor (John Richter).**

a). Changes made to the editor.

1. Dropped obsolete relationships: the editor no longer shows obsolete terms.
2. History now shows differently. You can now view the entire history of a term.
3. There is a new data adaptor that includes the relationship "develops from". But this only works with the database adaptor not flat files.

4. Saving to the database now works. The editor tracks all changes in your session, checks that you are working on the database and adds history tracking. From now on history will show everyone's saves. Can now query the history of a term-selecting a term highlights the edits for that term for the life of that term.
5. Conflict checking is working- save fails if a conflict is detected.
6. ID generation will be from the database. ID ranges will no longer be required.
7.All three ontologies are loaded when using the database adaptor.

b). In progress.

1. A plug-in for generating IDs other than GO IDs. For example, people using the editor to create anatomy ontologies will be able to have their own prefixes.
2. A gene association plug-in.
3. Update to require passwords for loading and saving to the database will be instituted.

c). Discussion/comments on the editor.

1. When importing from other flat files (e.g. anatomy) they can be parsed into the database by stripping terms. However, there may be reasons to save the original database identifiers –perhaps as synonyms.

2. While conflict checking works, top level node edits need to be announced to the group before starting.

3. When will the database be the real thing? Perhaps in about 3 weeks but, perhaps not.

**7. GO HTML Browser (Brad Mars):**

a.) Displays DAGs, gene associations, definitions. Searching by organism, term, etc… still some issues to resolve in the searching. BDGP has registered the domain name godatabase.org but this is not up yet.

b.) Some more work to be done on UI issues before release.

c ). Related to general software issues: Chris Mungall has been making progress with a BLAST server using GO annotated sequences from yeast.

**8. Website documentation issues:**

a). FTP site including CVS repository is moving to the outside of the firewall. This will allow anonymous read access to the CVS.

b). Updating to the FTP site. SGD has the only write access to the site. So in the event that Chris wants to dump the database XML files, Mike will grab the specified files from BDGP and load them.

c). DTD for the XML format needs to be moved to the website so that people can get to that. Currently the XML is updated once a month from Stanford to the database and then dumped out. This will move to nightly dumps once the DTD and XML is in place.

d) . Currently we are just exporting the tree but will also be exporting the associations (as above discussion on an XML format and appropriate DTD to be written).

e). The new format will allow people to specify and import subgraphs from the files.

f). A validation step will be added to the XML dump to make sure it is correct.

g). Move old documentation out of CVS but leave on the FTP site. Things to be archived include past minutes from meetings, old XML DTD files, old ontologies. Move the archive directory out of CVS but leave on the FTP site. Note that the abstracts/ directory is empty and could be removed.

h). As part of the rearrangement that Midori and Karen are undertaking , we can add a jobs page.

**9. GO expansion (specific issues and general concerns):**

a). Matt Berriman from the Pathogen Sequencing Unit at the Sanger Center raised some issues about how to expand the GO to include parasites. In particular the question was raised as how we can represent the components outside of the cell … moving from place to place. Alterations to component ontology need to be made to accommodate this.

A few basic ideas to deal with this:
1. add a host

```
%extracellular
  %host
    <host plasma membrane
    <host….
```

Could be another example of using cross products- with host and self as primary components. Can be implemented cross products that are used as needed or in the association files.
There is a problem with having host cell nucleus as instance of nucleus because this implies conflicting extra- and intracellular parentage. It was decided that as an extracellular term, host cell nucleus is not an instance of nucleus. A point arose from the discussion that GO terms are from the point of view of the organism being annotated. Therefore, host cell nucleus is not an instance of the intracelluar term nucleus, because nucleus can be regarded as meaning "a cell's own nucleus".

2. Combine terms with NCBI taxonomy ID. Probably wont work with parasites that have little host specificity.

3. Add a new type of relationship- belongs to (e.g. belongs to host).
Consensus is that the first approach (#1) is best. With the judicious choice of cross products.

If host is an instance of extracelluar it is not required to be an instance in nucleus.

b).DARPA- DAML language
This language has been selected by BioOntologies consortium to represent ontologies. Michael Ashburner will be working with this group to incorporate GO.

c). Other people/groups that have expressed an interest are Paul Kellam working on Herpes virus and a plant virus group.

d). General concerns:
1.A major issue is how big can we get and still be able to function as a group. Thirty people attended the Consortium meeting and this seemed like about as big as it can get. One solution is to break into smaller working groups that meet to deal with specific issues such as definitions or sections of the ontologies that need work. More about this in the organizational issues section.

2. Another issues is quality control-how to assure that the annotations make sense. What sort of annotation verifications are useful. Compugen has offered to help QA test and compare their results from literature and gene clustering with existing annotations. Possibly there could be a person in charge of monitoring gene associations and flagging inconsistencies. Another source of validation is the SGD four way comparison. Apparent inconsistencies can be flagged and sent to the appropriate group for review.

d). What is being annotated?
>Phenotype is clearly outside of the GO . There are no plans to expand to include phenotypes in the ontologies.
>Some groups, like SWISS-PROT, may be annotating to groups of protein products. Eventually, these should move out to the level of a transcript.
>In general we do not consider natural variants/polymorphisms, but it is really up to individual databases to decide on that.
>Splice variants, isozymes and other polymorphisms will be an issue for databases to deal with. Each variant should be annotated appropriately.

e). Expansion of the evidence codes. Generally agreed that more information needs to be provided so that it is clear how the annotation call was derived.
Case in point by Rolf Apweiler
For 26098 proteins
20757 from SWISS-PROT;5841 from Ensembl
Annotations include:

1448 mappings from EC
7696 SWISS-PROT keyword mapping
11025 InterPro mappings

1. 3022 unique GO terms that are all IEA. In general, there is not need to expand upon the IEA code but rather to have a more specific DB: reference for each type of annotation. So rather than dumping everything into IEA and giving a general reference, use IEA and the reference should explicitly state if the association was made from InterPro mapping or keyword mapping, etc. Send the references for the analysis as part of the annotation record. Each analysis method should have its own reference (e.g. GOFISH).

2. When is it an ISS vs. IEA call.. It is only ISS if a person has looked at it, in Swiss-Prot there is always a person looking at the sequence. Can we enforce a standard of reporting for sequence analysis methods? Probably this is unrealistic to attain.

The reference column of the association files will be used for providing more information on the different computational methods used for annotations. In addition, a set of descriptions of the methods used by each group will be put up on the GO website.

f). Integration of annotations from non-model organism databases.
SWISS-PROT and Compugen will provide first pass annotations to model organism genes to model organism databases to incorporate them and pass them on to the GO site. For non-model organisms, the associations will go directly to the GO site.

**10.Organizational issues**
a) . Anyone can participate in discussion. Suggestion of terms can and should come from anyone however write access to the ontologies should remain limited.

b). Who contributes to the annotations. The associations have to be a from a database.

c). What constitutes membership. In general requires that participating groups accept the principles of the GO and are willing to commit to ongoing development of the GO.
1.putting associations into the public domain.
2. contribution of financial support or data.
3.contribution of software tools.

d). Due to the size of the group (currently 10 members) it may be more efficient to have smaller working groups that meet for various reasons (such as an executive committee, or onotology working groups) in order to focus on specific action items.

To keep meeting at this rate (3-4 times a year) and to get things done, it may be helpful to change the structure of the meetings.

e). Should the GO site be a clearing house for information about other ontologies?

**11. More about cross products and anatomy.**

a). Cross products- David Hill expanded on the idea of cross products. We are in agreement to strip anatomy from process terms and proceed with cross products. For making cross products, it is important that the ontologies be orthogonal. We can expand the concept of cross products to many areas and it would be good to have a general tool for doing this that allows you to select specific nodes to create cross products (see d below). With respect to making anatomy ontologies for generating cross products with developmental process node in process ontology we need to first make orthogonal ontologies of anatomy and developmental stage, then take the appropriate cross products from these to make the cross product with development. It is essential to take the time component out from staging (e.g. days post-fertilization, post-germination are not useful as there is a lot of variation in how rapidly development occurs within a species).

An example:
[stage (organism specific, internal ID) X anatomy (organism specific, internal ID)] X [developmental process ( Go-generic, GO ID)] = GO ID.

Also, the cross-product of stage X anatomy will have an internal ID.

b). Anatomy browsing- Pavel from BDGP demonstrated a browser being developed at BDGP to display gene expression patterns and fly anatomy. Uses both images and text display.

c). Each organism database contributes their anatomy/ developmental stage ontologies and definitions to the GO and it will go into the CVS. Each group should be responsible (and responsive) for updates to their anatomy ontology.

d). Developing a tool to generate cross products. John Richter thinks he can adapt the editor to have this function. Realistically there will not be a tool till after October for generating cross products.

e). Report on papers citing or using the GO and activities related to GO.
1.MGI/RIKEN annotation paper is out.
2. Cathy Ball and SGD have their 4 way comparison paper in the works.
3. GO paper is out in Genome Research.
4.Matt Berriman -Parasites are GO in Trends in Parasitology.
5. David and Joel working on a paper about cross-products.
6.Courtland is doing a seminar for library sciences.
7. Michael will be doing a course in October.

8. TAIR review on plant data management includes small section on GO.
9. Postponed request from Annual Review of Genetics until next year.

f). Next meeting will be at the Chicago Omni Hotel and includes a users meeting. Regular meeting will be from the12th (new groups)-14th and users meeting on the 15th…