Gene Ontology Consortium Meeting

Divi Carina Hotel, St Croix, US Virgin Islands
January 25-26, 2002

Contents

Participant list
Progress Reports
Action Items from last meeting
Presentation: GO in UMLS: Jane Lomax
Content Issues
Database & Software
Annotation Issues
Miscellaneous

Documentation:
Appendix 1: Handouts accompanying progress reports
- A. GO Editorial Office, EBI
- B. FlyBase
- C. GOA @ EBI
- D. MGI
- E. GeneDB S. pombe (Sanger PSU)
- F. PSU (Sanger)
- G. SGD
- H. TAIR
- I. TIGR

Appendix 2: Action items from CSH May 2002
Appendix 3: Notes on J. Lomax presentation
Appendix 4: Assorted documents relevant to agenda items.
- A. Email from Tanya Berardini
- B. Email from Aubrey De Grey
- C. MGI Excessive granularity document
- D. MGI Negation document
- E. Documentation progress report (from Cath)

Appendix 5: Collected action items from this meeting

Participants

Michael Ashburner	FlyBase	Cambridge, UK
Daniel Barrell	EBI	Hinxton, UK
Matt Berriman	PSU(Sanger)	Hinxton, UK
Judith Blake	MGI	Bar Harbor, ME
Cath Brooksbank	EBI	Hinxton, UK
Evelyn Camon	EBI	Hinxton, UK
Tricia Dyck	DictyBase	Northwestern University, Chicago, IL
Kara Dollinski	SGD	Stanford, CA
Harold Drabkin	MGI	Bar Harbor, ME
Dianna Fisk	SGD	Stanford, CA
Becky Foulger	FlyBase	Cambridge, UK
Linda Hannick	TIGR	Rockville, MD
Midori Harris	EBI	Hinxton, UK
David Hill	MGI	Bar Harbor, ME
Eurie Hong	SGD	Stanford, CA
Amelia Ireland	EBI	Hinxton, UK
Jane Lomax	EBI	Hinxton, UK
Brad Marshall	BDGP	Berkeley, CA
Suparna Mundodi	TAIR	Carnegie Inst., Stanford, CA
Chris Mungall	BDGP	Berkeley, CA
Sue Rhee	TAIR	Carnegie Inst., Stanford, CA
John Richter	BDGP	Berkeley, CA
Valerie Wood	GeneDB S. pombe (Sanger PSU)	Hinxton, UK

Progress Reports

For full reports, see Appendix 1.

GO Editorial Office, EBI
- over 600 new terms added; 70% of terms now have definitions
- every GO synonym examined and a relationship to the term name assigned (as part of UMLS project)
- comments added to all obsolete terms
- SourceForge item notification script now up and running

DictyBase
- public beta of annotations is viewable and will be added to the GO repository after checking
- medical ontology has been developed and should be available soon

FlyBase
- 27,056 GO annotations now in FlyBase
- Swiss-Prot GO annotations continuing - these include annotations for non-D. melanogaster genes.
- most recent re-annotation of the Drosophila genome (release 3) is almost complete
- definitions added to a number of fly specific process terms

GOA @ EBI
- 6 GOA-SPTR releases, 8 GOA-Human releases
- GOA dataset to be enhanced by mappings from Swiss Institute of Bioinformatics
- GOA cross-referenced directly in the EMBL nucleotide sequence database
- QuickGO browser updated

MGI
- annotations added at a steady rate; 41,000+ annotations to 9032 genes
- continued development of phenotype ontology; expected to be made public by mid-February
- RIKEN data has been loaded into the database

GeneDB S. pombe (Sanger PSU)
- total of 15,029 GO term assignments now made to process and component terms
- extensive overhaul of configuration files to give constant refinement of associations

PSU (Sanger)
- full manually curated GO annotation of malaria finished
- joint curation with TIGR of Trypanosoma brucei continues
- annotation of Aspergillus fumigatus and Theileria annulata genomes to come

SGD
- two new software tools: GO Term Finder and GO Tree View
- every ORF at SGD has a function and process term annotation
- every named ORF has a complete set of GO annotations

TAIR
- GO terms being added to the ontologies with definitions
- plant GO-slim developed and submitted
- aim to annotate all studied Arabidopsis genes to all three GO ontologies

TIGR

- T. brucei chromosomes 4 and 6, rice and Aspergillus fumigatus are in the works
- Shewanella association file recently submitted
- several bacterial genomes awaiting publication

ACTION ITEM: TAIR to update MetaCyc2GO mappings.

Action Items from last meeting

See Appendix 2 for full details. Action items arising from this were:

ACTION ITEM: John. 7 from last time [add term deletion feature to DAG-Edit].

ACTION ITEM: Brad. 10 and 11 [adding more information about GO curators to website/database] outstanding.

ACTION ITEM: Come up with system for notifying developers of format changes.

ACTION ITEM: Add "contributed by" column.

GO in UMLS : Jane Lomax

See Appendix 3 for the full presentation.

Progress report: GO has not yet been released with UMLS Metathesaurus, but substantial progress has been made. There has been a successful insertion of the molecular function ontology, with cellular component and biological process soon to follow. There are two major issues created for GO; how to handle GO 'synonyms', and ambiguity in GO term names. These issues are discussed later in the meeting.

Content Issues

Synonyms: distinguishing exact synonyms from related terms
- how many types to distinguish?
- how to store/represent (implications for tools)?

There was some discussion regarding the synonym types. In particular, whether a synonym with the "broader than" relationship to the main term reflects a missing parent or relationship in the tree, and also the number of relationships we need - do we need finer distinctions than true synonym vs related term? It was concluded that we would keep all the existing types of synonyms (exact, broader, narrower, related to, undefined) and the hierarchy of synonym types would be as follows:
related to
[i] exact
[i] broader
[i] narrower
[i] undefined

ACTION ITEM: Curators. When adding new synonyms, track which type they are. If they are 'broader than' or 'narrower than', consider whether it calls for a new term.

ACTION ITEM: Jane. Circulate synonym list again.

ACTION ITEM: BDGP. Look into rules that could be worked into DAG-Edit to make synonym maintenance easier.

GO/UMLS component term merge problems

The problem stems from ambiguity in term names. The term string "xxx complex" in GO refers to a cellular location, but the same string in UMLS usually refers to a protein entity and would be assigned the semantic type 'amino acid, peptide or protein'. The question is, does the GO cellular component term mean the same as the UMLS concept? If it doesn't, and a new concept would have to be created, what semantic type should we assign it, and what relationship would need to be created between these new and existing concepts?

It was agreed that the GO 'xxx component' cellular component terms were different in meaning to the existing 'xxx complex' concepts in UMLS, and GO term names should not be changed to fit with UMLS. It was decided that Jane should discuss possible solutions with UMLS people; possibly modify some GO term names in UMLS only (by adding 'location'?) or see whether UMLS can help come up with a solution in their system, and to keep consortium informed of progress.

The consensus was that all cellular component terms should be in concepts with the semantic type 'cell component' (never part of a concept with the semantic type 'amino acid, peptide or protein') and that the relationship between the new (with GO term) and existing concepts should be something broad, like 'related to'.

ACTION ITEM: Jane. Discuss this with UMLS and fill us in on the results.

Cellular processes: questions to be resolved before the cellular process reorganization is committed

See Appendix 4A for the email from Tanya Berardini containing the questions.

- Cellular differentiation vs cell fate commitment and cell type development vs cell type differentiation

David Hill outlined a suggestion: cell differentiation can be broken down into the following steps; cell fate commitment where a cell senses its location and begins to specialize, but can still switch types, cell type determination where a cell switches irreversibly to a specific type and cell development where a cell physiologically matures into its type. Should we use these divisions in GO? The group agreed that we should.

Conclusion: Cell differentiation and its children will have the following structure:
cellular process
[i] cell differentiation
  [p] cell fate commitment (exact synonym: cell fate specification)
  [p] cell fate determination
  [p] cell development (exact synonyms: cell morphogenesis, cell maturation)

- Response to endogenous stimulus and response to exogenous stimulus

Cellular response and organismal responses are usually linked; we would like to capture relationship but don't want to violate true paths (eg. for unicellular orgs). This means being very careful with parentage. Cue a big discussion of where to put the unicellular/multicellular split. A working solution was proposed: make the split as far below 'physiological process ; GO:0007582' as possible, and as and when needed, rather than splitting right below physiological processes. We will revisit this to see how the solution has worked. Leaving the "response to xxx" terms under cell communication is fine.

The group agreed that it was always important to keep annotation in mind when making these changes, and reaffirmed the need to keep GO process terms covering multicellular processes, as they are needed for annotation in many species and help in the development of orthogonal ontologies.

ACTION ITEM: David and Tanya. When splitting out multicellular vs unicellular processes, make the split as far below 'physiological process ; GO:0007582' as possible, and as and when needed, rather than splitting right below physiological processes.

Grouping terms in the function ontology

Prompted by Karen's email on G-nucleotide release factors and the related items RNA polymerase and hydrogen-translocating ATPases

The function ontology contains grouping terms that reflect process or component info (eg. DNA repair protein; membrane-associated functions). This cross-contamination is useful for helping curators find terms but is not consistent with the guidelines set out for function terms. One approach would be to make relationships between the function and component or process ontologies and remove the grouping terms. This would require VERY careful curation as some functions act in

many processes. A better solution would be to expand the toolset available to curators, eg. Fritz Roth's statistical links and concurrent assignment tools.

The conclusions were that no hard-coded links will be made between the ontologies and instead research would continue into tools to make statistical links.

ACTION ITEM: GO editorial team (and others). Start removing grouping terms slowly and carefully with all the usual communications. If obsoleting a term, ensure the corresponding process or component exists.

Should functions (particularly enzyme functions) be differentiated on the basis of environment?

1. pH-specific enzymes: Example given was GO:0030230 and GO:0030231, differentiated on the basis of the pH at which they act.
Conclusion: different EC numbers - keep both terms; same EC numbers - obsolete the pH-specific examples and use the parent term.

2. Hydrogenases: Example given was GO:0008901 and its children GO:0016948 - GO:0016951. They have the same EC number but different metal ions associated with them. This could be solved in the same way as protein binding - at the annotation stage, use a chemical ontology and use the extra column to note the metal. Alternatively, we could use multiple parents and/or annotate to separate terms (eg. hydrogenase, iron binding). The issue was not resolved after discussion and will probably be left until we have software to implement the new column.

Should we add 'activity' to function term strings?
- if so, do we change the main term string or add 'related terms'?

Two main arguments for this: first, it reduces the ambiguity of the term name, therefore helping when GO is included in other systems (specifically UMLS), and second, it will reduce user confusion. All agreed this was a timely step.

ACTION ITEM: Jane. Add activity to function term strings.

How to represent membrane proteins
- whether to have 'integral [to] membrane', what wording
- whether to add children (e.g. for type I, II, III, IV transmembrane)

In the component ontology, we used to have 'intergral membrane protein' plus children which was problematic because it didn't refer to a location, rather a relationship between a membrane protein and a membrane. The wording was recently changed to 'integral to membrane'; did we want to keep this for the long term or find some other solution? The other issue, brought up by Evelyn, was whether to add more granular child terms for the different types of transmembrane protein, as this would help with Swiss-Prot/GO mappings. This idea was rejected because these are types of protein and not locations.

Conclusion: Keep the membrane terms as they are now (integral and peripheral); don't add the children as they don't reflect a location.

Should the 'host' term be used for viral cellular component terms?

The term 'host' was originally created for describing the cellular component of single-celled parasites infecting a host cell, so it was placed under 'extracellular'. A problem arose when trying to add the new viral terms, because viruses aren't cells, so the host cell environment is not extracellular. Various options were discussed, including moving 'host' out from under 'extracellular', but it was felt that the best option was to simply extend the definition of 'extracellular' so that it could be applied to organisms that aren't technically cells. A comment would also be added explaining why this was done.

ACTION ITEM: GO editorial team. Define extracellular to include outside a virus particle, then use host terms as parents for the appropriate virus cell component terms.

How should we handle component terms that can be both intracellular and extracellular?

Some complexes can be intra- or extracellular; the example given was 'immunoglobulin complex ; GO:0019814' which can be either membrane bound or circulating, so there are two is_a child terms, 'immunoglobulin, circulating ; GO:??' and 'immunoglobulin, membrane bound ; GO:00??'. The problem comes with the placement of the parent term, the generic 'immunoglobulin complex', which might be used when you know that a gene product is a component of an immunoglobulin molecule, but not know whether it is membrane bound or cirulating. At the moment the term is placed directly under 'cellular component', but it's going to end up a pretty long list!

After some discussion, during which we considered whether we needed a generic term at all, it was felt that the most appropriate place for such terms is directly under cellular component where we currently have them.

ACTION ITEM: GO editorial team. Go through the enzyme complexes (see also SF entry 535294) and where applicable, make a general parent directly under 'cellular component' with children in specific locations.

Term grammar (for use in automated construction of sentences describing gene products)

See Appendix 4B for the email from Aubrey de Grey

We are willing to alter the term grammar to suit Aubrey's needs as long as:
A: Aubrey sends terms so we don't have too much work to do!
B: we check carefully to make sure any changes won't wreck terms for biologists searching or curators annotating

ACTION ITEM: GO editorial team to get list from Aubrey and evaluate; adjust terms as needed.

Revisit 'catalyst' and 'regulator' part-of children of some enzymatic activity terms

Several enzymes are split into a catalyst and a regulator function. This item questioned the need for these terms as they sound like enzyme components rather than functions. After discussion, it was decided that they should be left as they are to allow maximal information about protein function to be captured.

Revisit the "Round Table Discussion" on how to represent synthesis/binding/etc. of individual proteins

See Appendix 4C for the MGI excessive granularity document.

The problem is basically that GO cannot allow gene product names inside GO terms because of the rampant proliferation of terms that this generates, however, it is still useful to be able to annotate to this level of granularity. For instance, to able to state that a gene product IL18_HUMAN is involved in 'interleukin-13 biosynthesis'.

The solution proposed by Chris was as follows; some GO terms would have 'slots', which would be filled in the gene_associations file. For instance, 'biosynthesis' would have a 'slot' named 'synthesizes'. The GO term 'interleukin-13 biosynthesis' would therefore not exist, and instead, the annotation for IL18_HUMAN would include an entry to GO term 'cytokine biosynthesis ; GO:0042089' or just plain 'biosynthesis ; GO:0009058'; this entry/line would also have a column for 'slot', which would read "synthesizes(interleukin-13)". Interleukin-13 could be replaced with an identifier from a product/family/physical-entity ontology.

The proposition is described in more detail at http://www.fruitfly.org/~cjm/slots.html

The practical implications were discussed; there is a need for ontologies to cite in the slot values, for example, a chemical ontology and a protein family ontology. A few exist and more will be available in about a year. This will also require a rethink of annotation practice, and some new tools. Existing annotations would of course have to be retrofitted, but the bulk of this could be automated. Of great importance is considering our users, any changes need to be announced well in advance. In addition, would we change the front-end appearance of tools, e.g. AmiGO, or keep these changes behind the scenes? One issue is that using the slots effectively creates GO terms that are cross-products, but do we instantiate these products - i.e. give them GO IDs? For instance, if we were to instantiate all the terms generated by the cross product between 'synthesis' and a product/molecule/chemical ontology we would have actual GO IDs:

GO:9000001 IL-1 biosynthesis
GO:9000002 IL-2 biosynthesis
GO:9000003 IL-3 biosynthesis
GO:9000004 IL-4 biosynthesis
GO:9000005 IL-5 biosynthesis

The disadvantage is that any time the orthogonal ontology of products is changed, GO has to be changed (either manually or automatically) to reflect this. For example, if IL1 was split into IL-1a, IL-1b we would need IL-1{a,b} {biosynthesis, receptor} etc in GO.

With the 'slots' approach there would be no GO ID for "IL-8 biosynthesis". Curators could still annotated genes as "IL-8 biosynthesis" by dynamically combining the terms using slots but the disadvantage is that there would not be a single GO ID they could quote in a paper etc.

ACTION ITEM: Announce on the website that we'll implement this solution at some future date (no date set but will be 6+ months from now). Assemble a group (MA, Chris, David) to work on the implementation.

Interest Groups

Interest groups and areas have been extensively examined or claimed already, the problem is, how to ensure that the interest group is informed when changes are made to that part of the ontology? We could have interest groups listed e.g. in SourceForge, or on our webpage, perhaps with a list of GO_Slim terms defining the area of interest alongside. Anyone making changes to these areas would then have to inform these groups first, then the onus would be on these groups to pipe up if they had a problem!

ACTION ITEM: Midori to put up interest groups on web page. Everybody to send group ideas & which they volunteer for. See if it works or if we need further formalization by putting groups in SourceForge.

Annotation

Annotation of disease genes

Annotations of genes implicated in disease to be submitted by Nat Goodman. These should be fine as long as he doesn't annotate actual disease processes, i.e. he must only annotate the normal functions of genes implicated in disease.

Consistency and quality control
- Suggestion from Evelyn: a set of "standard annotations" for common proteins.

Evelyn has seen different terms assigned to "common" proteins; is this a QC problem or does it differ between organisms and what has been studied and what experiments have been done? How do you define "common proteins"?

Conclusion: Annotations are the responsibility of individual databases. Differences often reflect the state of experimentation. Evelyn has unique perspective for spotting inconsistencies, because SWISS-PROT includes annotations from all organisms. She should keep communicating problems to the individual databases.

Negation

See Appendix 4D for MGI's handout.

Conclusion: The best solution in the long term is to use Chris's slots model; in the meanwhile, muddle through somehow - each group can decide what works best for them.

Database and Software

DAG-Edit & GOET

One line of GOET work has stopped, but GOET overall goes on. John is back working on DAG-Edit. :-)

New DAG-Edit features (full list appears in the release notes of the latest version):

- - Search tool remembers last 10 searches on each field
- - Configuration plugin allows users to show undefined terms in gray
- - Changed flat file format to support multi-character types
- - The available relationship types are now defined per-session, instead of per-adapter
- - Created a Relationship Type Manager plugin that allows a user to define which types are available in a session
- - Dbxrefs now have an editable description (however, the flat file format cannot store these descriptions)
- - An arbitrary number of files can now be read in at one time (instead of just 3)
- - File read history now stores groups of files, not one file at a time

John would like switch over to the new flat file format. This should be announced on the proposed webpage for forthcoming software/data format changes, as well as on the GO site in SourceForge. Users should be given adequate time to switch over; John suggests allowing two months after the announcement has gone up.

The new format allows relationship symbols & types defined in headers, and multi-character relationship types are possible, as well as dbxref comments in the flat file. It also has a reduced file size due to non-redundant display of parentage.

Other planned features for DAG-Edit include:

- - multiple terms viewable in gene product plug in (only one can be viewed at a time at the moment)
- - option to have a "delete" button to move terms to obsolete
- - plug-in for cross-products
- - spellcheck function to use with the dictionary file

AmiGO

Brad reported that the AmiGO GOst BLAST server is now live. He also reported that an AmiGO software upgrade is coming soon. Brad is interested in feedback from the community that uses GO on what data and tools they use and how they use them.

ACTION ITEM: Construct and post a user survey covering tools, AmiGO, etc.. Send question ideas to Amelia Ireland. It will be sent out to GO-Friends and data collected in time for the grant application.

Database

There isn't much change to report on the database. Chris and Dave Emmert (Harvard) are developing CHADO, a postgres database. It will be more capable of holding different ontologies; it is expected that FlyBase and GMOD will use it and it will probably subsume the GO database.

Database Updates (Chris Mungall)

Chris says that automated database update are taking place approximately once a month. He has a script which creates 4 downloads: terms; terms and annotations; terms, annotations and sequences; terms, annotations without IEA and sequences (for AmiGO). The script takes takes approximately two days to run. It was suggested that there should be a daily update of the database terms and structure to prevent the lag seen between the addition of the new terms and their appearance in AmiGO.

ACTION ITEM: Chris. Suggestion: a daily release of a separate database containing just terms without annotations. The whole database should be updated every month. AmiGO would have the option to view the up-to-date term set with no associations.

Chris has scripts to map gene association files to GO-slim terms; it uses 'bucket' terms such as "other enzyme" which are given temporary GO-slim IDs.

ACTION ITEM: Chris. Make use of parents rather than bucket terms to avoid confusion due to transient IDs.

Brad clarified the AmiGO pie chart maker behaviour and accepted suggestions for new features.

ACTION ITEM: Brad. Investigate piping GO-Slim mapping results to the AmiGO pie chart maker.

ACTION ITEM: Brad. Add the ability to dump AmiGO pie chart data as a flat file containing GO ID, term name and the number of gene products.

Miscellaneous

GO.bib file

During the updating of the documentation, Cath discovered the GO.bib file and asked who uses and maintains it and whether some guidelines could be drawn up for its content and usage. It was concluded that no one uses this document (let alone maintains it!) and it could be removed from the GO documentation.

GOBO

After the success of the Standards and Ontologies for Functional Genomics (SOFG) conference, Helen Parkinson (EBI) has had requests for an ontology site hosted at the EBI or at sofg.org. Michael Ashburner will talk to Helen and Chris Stoeckert about this.

Documentation

See Appendix 5E for Cath's progress report.

Cath has made significant progress in her work on the documentation. Unfortunately, Cath is no longer part of the GO team at the EBI, but she was able to do the work in her new role as part of the Outreach team. She has reorganized, rewritten and updated the documentation to make it clearer and easier for users to find the information they are looking for; to this end, she has split the information into sections relating to different GO users. There were several action items relating to the documentation:

ACTION ITEM: Member databases. Each database should send annotation FAQs from their existing documentation to Cath for inclusion in GO FAQ. GO FAQ will have general annotation FAQs and then specific FAQs from each database and from the EBI.

ACTION ITEM: Everyone . Read over the new documentation (especially the style guide) and send any suggestions to Cath. This is available at http://www.ebi.ac.uk/~cath/

ACTION ITEM: Cath. The changeover to the new documentation will occur on 15 March.

ACTION ITEM: Cath. Update the synonym section of format guide to accommodate the decisions made at this meeting.

ACTION ITEM: Chris. Provide some documentation on the mySQL database.

ACTION ITEM: Jane and John. Update the DAG-Edit user guide.

Grant Proposal

Judy reviewed the schedule and plan for the upcoming competitive grant renewal for the GO Consortium. We will submit our proposal to the NHGRI on March 1. We will ask for continued support for the development of the ontologies, now including the Sequence Ontology for sequence features. We will ask for continued support for the annotation of genomes and gene products to the

GO by the model organism databases and Swiss-Prot. We will ask for continued support for a community database resource which includes open access to the ontologies, the annotations to the GO, and other resources and tools. Some new aspects of the project are that we will continue to work to provide the ontologies in DAML+OIL, and will provide support for pilot projects that investigate or interact with the GO in new ways.

Next Meeting

Host: TIGR, June 3 - 4 (no users meeting).
Minutes: BDGP

GO Editorial Office – EMBL-EBI

Progress Report

- Summary of changes to ontology since last meeting:

	Number before last meeting	Number now	Number new	%
component	1108	1158	50	4.5% increase
process	6415	6889	474	7.4% increase
function	5231	5322	81	1.5% increase
total	12754	13373	619	4.9% increase
defs	7730	9368	1638	21.2% increase now 70% defined

Other news highlights

- Around 100 new viral terms added in collaboration with Ria Holtzerland from University College London.
- Comments added to *all* obsolete terms, giving a reason for obsoletion (wherever possible) and alternative terms (if available).
- Every GO synonym examined and a relationship to the main term assigned (as part of the UMLS project).
- Script written in response to Evelyn's request for a monthly list of obsoleted terms with the suggested alternatives. Monthly digests available soon.
- New ontology checking procedures added, catching errors such as redundant relationships
    - term losses
    - incorrectly formatted EC, ISBN or TC dbxrefs
    - definition formatting errors
- Midori's script for a daily digest of SourceForge requests up and running.

Appendix 1B. FlyBase

FlyBase Gene Ontology Progress Report. January 2003.

1. CURRENT GO ANNOTATIONS IN FLYBASE
(Stats taken on January 14th 2003)

Total genes annotated with at least 1 GO term:	7,387	
Total number of process terms:		9,238
Total number of function terms:	11,282	
Total number of component terms:	6,536	

Total number of GO annotations in FlyBase: 27,056
(includes 113 annotations with the IEA evidence code)

2. ANNOTATION

SWISS-PROT annotation
Eleanor Whitfield at SWISS-PROT is continuing to send GO annotations of new SWISS-PROT records and SWISS-PROT records updated from SPTrEMBL that are linked to a Drosophila gene. These are incorporated into our files using the evidence codes and references Eleanor provides. They are internally tagged so that her annotations can be traced. These are the first GO annotations we have for non-melanogaster Drosophila genes. We now have at least one GO-annotated gene for the following non-melanogaster species.

D.auraria (1)
D.erecta (1)
D.funebris (1)
D.hydei (1)
D.mauritiana (5)
D.miranda (1)
D.orena (2)
D.persimilis (1)
D.pseudoobscura bogotana (1)
D.pseudoobscura pseudoobscura (3)
D.sechellia (5)
D.simulans (6)
D.subobscura (1)
D.takahashii (1)
D.teissieri (3)
D.virilis (9)
D.yakuba (4)

(Numbers in brackets refer to the number of genes in these species that are annotated with one or more GO terms).

Literature Curation
FlyBase curators are continuing to add GO terms from literature curation of primary papers (11 journal titles are curated on a regular basis and a further 30 are curated when time permits) and personal communications.

Since FlyBase was founded 6 years before the GO project began, there is a lot of data in FlyBase that could be annotated as GO terms. To rectify this, recent reviews have been curated to increase the number of process GO terms in FlyBase.

Electronic Annotation
Following on from previous data sets, in October 2002 Fritz Roth provided FlyBase (and SGD) with 50 predicted GO terms based on existing GO terms and patterns of annotation. The FlyBase predictions were manually assessed, validated, and Fritz submitted a paper describing the project.

Sequence Curation
The most recent re-annotation of the Drosophila genome (Release 3) is almost complete. This has resulted in changes to a number of gene models, namely:

i. Gene splits: one gene being split into two or more genes.
ii. Gene merges: two or more genes being merged into one gene.
iii. Gene splerges: a mixture of splits and merges of one or more genes into one or more new genes.
iv. Change in coding sequence of a gene.
v. New gene predictions.

For the gene splits, merges and splerges, the existing GO data from under the genes has been temporarily removed. The Release 3 gene models are currently being analysed and GO data assigned accordingly, based principally on sequence similarity to gene products in other organisms. Following this, new genes and genes where coding sequence has changed will be analyzed for potential GO terms. The majority of time over the next month will be spent on Release 3 GO annotation.

3. DEVELOPMENT OF THE ONTOLOGIES

Definitions have been added to a number of fly-specific process terms, e.g. the events involved in dorsal closure and insect tracheal morphogenesis. Once the Release 3 GO annotation task is complete, a principal project will be to revise and provide definitions for the fly-specific development terms in the process ontology.

Rebecca Foulger and Michael Ashburner

e-mail: r.foulger@gen.cam.ac.uk
m.ashburner@gen.cam.ac.uk

http://flybase.bio.indiana.edu/
http://fly.ebi.ac.uk (UK mirror)

GENE ONTOLOGY ANNOTATION (GOA) PROJECT, EBI.

ST. CROIX GO CONSORTIUM REPORT 25-JAN-2003

Contacts:

Rolf Apweiler (apweiler@ebi.ac.uk), Evelyn Camon (camon@ebi.ac.uk), Daniel Barrell (dbarrell@ebi.ac.uk), goa@ebi.ac.uk. URL:http://www.ebi.ac.uk/GOA

Contents:

1. Current status
2. New Cross References
3. New SWISS-PROT Mappings/Updates
4. Data Integration
5. QuickGO Browser Update
6. Databases using GO/GOA at EBI.
7. GOA Publications
8. Future Activities

1. Current Status:

5 releases GOA-SPTR, >2.56 mill associations, >549,000 SPTR entries, 49701 species, 34439 Pubmed References (10937 distinct),~64% GO Coverage of SWISS-PROT and TrEMBL.

7 releases GOA-Human, 74872 associations, 18554 SPTR entries, 24188 Pubmed References (9756 distinct).
Proteome Inc. obsoletes removed/replaced with parent term. Proteome Inc. evidence codes now replaced with GO evidence codes. Pathogenesis terms still need to be removed/manually revised.
Human GO annotation marathon expected Summer 2003 (see point 8).
GOA is ahead of schedule on all grants.

2. New Cross References

Dec 15 2002, GOA has been cross-referenced directly in the EMBL Nucleotide Sequence Database. 734286 coding sequences (CDS features) in EMBL now have a cross-reference to GOA e.g. /db_xref="GOA:P01100", and these are hyperlinked in the EBI SRS server to the GO annotation displayed in our QuickGO browser.

Cross references to GO terms using GOA data is provided in XML version of TrEMBL. Aim by Feb/Mar, 2003 to cross reference GO in flatfile version of both SWISS-PROT and TrEMBL,
(proposed format):

DR GO; GO:0004984; IDA; Function. Or..
DR GO; GO:0004984; IDA; F.

Or in the CC (Comment lines as:)
CC -!- GO FUNCTION: GABA-A receptor (GO:0004890)
CC -!- GO COMPONENT: centromere (GO:0005698)

Discussion on whether GO Term name should be included, considered too long for flatfile but may be possible to display fragmented term or reserve term display for NiceProt view.

3. New SWISS-PROT Mappings/Updates

In 2003, the GOA dataset will be further enhanced with new GO mappings from SWISS INSTITUTE OF BIOINFORMATICS (SIB) for SWISS-PROT's subcellular location as well as HAMAP (High-quality Automated and Manual annotation op microbrial Proteomes, (contains a collection of manually curated microbial protein families)) and PROSITE (database of protein families and domains). HAMAP curators will be trained in GO annotation end of February 2003. We have some new Talisman tools to create at EBI to help SIB mapping management.

SIB will also check InterPro2go mappings and feedback to Nicky Mulder at EBI. There was a concern that InterPro2go can sometimes overpredict GO terms, (i.e too specific). This is because InterPro curators currently don't get notification when new SPTR entry is integrated into old InterPro entry with GO annotation. Nicky will try to resolve this matter.

There has been a request to suppress InterPro2GO annotation to unknown function/process or component, the request has been rejected,
Nicky thinks they may be useful (opinions please).

Spkw2go has been updated Jan 2003.

We are still working closely with PIR to help their keyword mappings as part of UniProt Consortium.

4. Data Integration.

GOA hopes to start integration of GO annotation from other Consortium members by February 2003. This is now possible because of the ability to acknowledge the source database in an extra column in the gene association file. Non-GO Consortium members

from specialised human databases wish to submit their GO annotation to us. This will be permitted on case by case basis to some specialist database groups as long as they follow the GO Annotation Guidelines. Individual scientists should continue to submit updates to GOA via goa@ebi.ac.uk mailing list.

5. QuickGO Browser Update.

The QuickGO browser has been updated but the new version is not released yet. The new version will be more stable. New functions: Can display all or only manual curated entries. GO ontology can be viewed as denormalised view or tree view. Common Concurrent assignments have been updated. GO Comments will also be viewable.

6. Databases using GO/GOA at EBI.

EMBL, SWISS-PROT, TrEMBL, InterPro, Ensembl, AltSplice DB, IntAct, IntEnz, ArrayExpress, MSD, (Resid).

7. GOA Publications

There is already 1 InterPro publication explaining GO:

Biswas M., O'Rourke J.F.,Camon E.,Karavidopoulou Y., Kersey P., Kriventseva E., Mittard, V., Mulder N., Phan I., and Servant F. Applications of InterPro in protein annotation and genome analysis. *Brief Bioinform.*3:285-295(2002).

2 further GOA publications are *in press.*
Comparative Functional Genomics (ESF Ontology for Biology).
Genome Research.

8. Future Activities.

Human GO annotation marathon expected Summer 2003 as part of collaboration with Alphonso Valencia (Coordinator of the Spanish Network on Bioinformatics). This is part of data mining experiment to find tools that can accurately predict GO terms. SWISS-PROT at EBI will produce a highly annotated set of human GO annotations for comparison with various data mining techniques used across various institutes. During the competition there will be a delay in releasing new Human GO annotation so that the bioinformaticians can't cheat!

SRS retrieval of GO terms using GOA needs attention to allow complex querying, probably will not display GOA in GO 'gene association' file format. Complex querying will be possible when GO is directly cross- referenced in SWISS-PROT and TrEMBL flatfiles.

Appendix 1D. MGI

General:

We continue to focus on extending our goal to have annotation for all genes in the database. This includes both adding annotation to genes currently without any annotation, and replacing annotations that were "fished" from text records with literature based annotation. We appear to be adding annotations at a constant rate (Figure 1). Progress since the last meeting is summarized below:

MGI GO STATS as of January, 2003.

Annotation Type	**23-Jan-03**	**27-Aug-02**	**Change**	**% Change**
Total Genes annotated:[1]	9032	8576	456	5.3
Total Hand Annotation				
# of Genes	3501	2646	855	32.3
Orthology:	27	24	3	12.5
"IEA"				
SwissProt to GO	6114	6123[2]	-9	-0.15
Interpro to GO	3528	3529	-1	-0.03
EC to GO	653	658	-5	-0.75
MLC Scan	40	40	0	0
GO Fish	2199	2228	-29[3]	-1.3

Beyond GO

The phenotype ontology continues to be developed with the aid of the DAG-Editor[4]. The expanded Phenotypes Classification is scheduled to be publicly available by mid-February (see Figure 2).

[1] Number of genes with at least ONE GO term of any kind.

[2] Decreased due to movement to obsolete. This also holds for Interpro and EC to GO

[3] This figure has decreased due to our ongoing efforts to replace these with literature based annotation..

[4] Cynthia Smith, Cathleen Lutz, Carroll Goldsmith, Teresa Chu, and Alan P. Davis

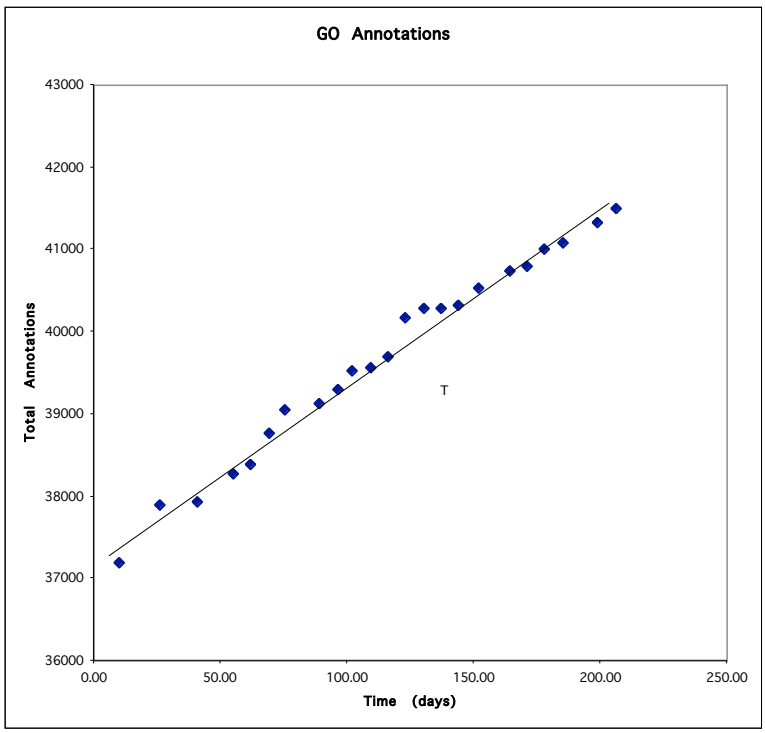


Figure 1 GO Annotation Increase

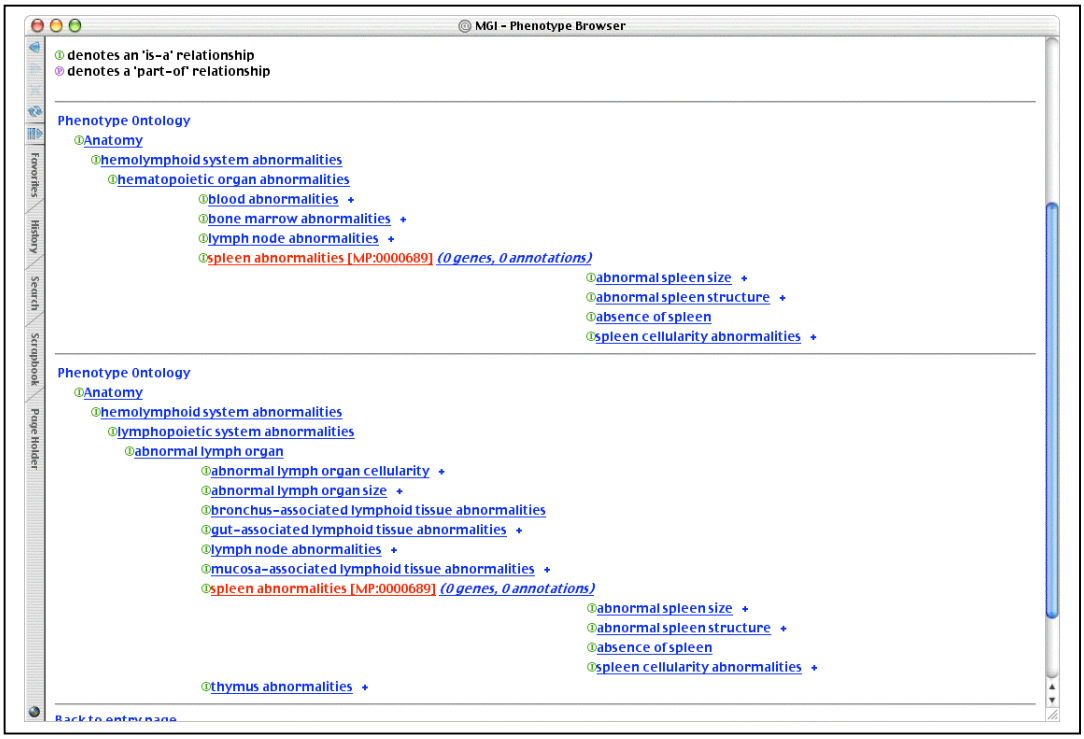


Figure 2 Phenotype Classification Browser

GO meeting report *Schizosaccharomyces pombe*

Brief Sequencing Status

The 3 chromosomes of the 13.8 Mb genome of *Schizosaccharomyces pombe* are currently in 8 contigs. Attempts to capture the remaining telomeric and centromeric gaps are still in progress.

The genome contains 4966 ORFs including mitochondrial genes and transposons and approximately 600 identified or predicted RNAs.

Annotation Methods

Manual Curation

Curated descriptions are attached to each gene using structured syntax. These are either PubMed supported, or inferred from similarity.

e.g. PubMed supported

Name	ecm5
Systematic Name	SPBC83.07
Status	experimentally characterised or published
Description	Lid2 complex PMID:12488447 2 others implicated in transcriptional regulation PMID:12488447 interacts physically with SET1 complex PMID:12488447 implicated in regulation of chromatin remodelling PMID:12488447 zinc finger protein 190 others similar to S. cerevisiae YER169W

e.g. Inferred from similarity

Systematic Name	SPBC83.07
Status	role inferred from homology MC transporter of unknown specificity 10 others similar to S. cerevisiae YMR241W

Structured syntax is constantly under revision and provides:

Grouping of similar terms analogous to GO terms but also encompasses interaction, domain, similarity, species distribution, post translational modification, status etc. .

An additional data entry point, and facility for query result verification. For instance zinc finger proteins, zf C3HC4 type are identified from Pfam, SMART, published literature, or inferred from context and are grouped together by use of structured syntax.

Improved curation; similar descriptions are gradually grouped and global changes can be implemented together.

Ease of parsing into the relevant tables of the relational database data are stored in flat files until the relational database is implemented .

GO Term Assignment

Schizosaccharomyces pombe GO assignments are made semi automatically. Curated descriptions see Manual Curation above are compared to configuration files containing sets of 'curated keywords' which are always associated with a particular GO term.

e.g. Configuration file rRNA processing partial

RRNA METHYLTRANSFERASE
RRNA LARGE SUBUNIT METHYLTRANSFERASE
PROCESSOME
PRE RRNA CLEAVAGE
PSEUDOURIDINE SYNTHASE
RRNA PSEUDOURIDINYLATION
RRNA BIOGENESIS
RRNA MATURATION
RRNA PROCESSING
RRNA PSEUDOURIDINYLATION
EXOSOME

Use of standard syntax in annotations and configuration files results in constant refinement of associations.

Caveats:

The automated aspect means that for the present all associations are ISS, even if experimental data is available. However, the source of the association is always traceable in GeneDB via the manual curations.

Annotations are often NOT made to the most detailed level of the ontology because only a restricted subclass of terms are used. However, the terms used are extended by implementing available child terms when the number of assignments to a node reaches 100.

Associations are often made to parent AND child terms, not only the most specific term.

For example, a glutamate transporter would be annotated to each of these 3 nodes:
GO:0006519 Amino Acid and derivative Metabolism 184
GO:0006520 Amino Acid Metabolism 163
GO:0006536 Glutamate Metabolism curated 26

This is implemented for practical purposes of data retrieval. Future releases will consider either purging all derived parents from the associations files, and/or purging configuration files of more specific terms.

Annotation Status

	Oct 2002	Jan 2003
Curated descriptions	15391	16705
Curated descriptions PMED supported	1190	1940
GO process assignments	10441	10300
GO component assignments	4757	4729
Total GO assignments	15198	15029
GO process terms used	135	129
GO component terms used	?	74
Genes with at least 1 process term	3412	3507
Genes with at least 1 component term	?	2010
No. with curated S. cerevisiae orthologs	2373	3223

New papers are curated as published. Old papers are curated *ad hoc*.

The small drop in GO assignments since last meeting is due to an extensive overhaul of the configuration files

Obsolete GO term assignments have been fixed

Future Aims

Add GO slim function associations

Finalize overhaul of the configuration files

Documentation for sequence analysis and annotation including criteria for similarity assignments

Appendix 1F. PSU (Sanger)

The full manually curated GO annotation of malaria is now in GO cvs and is ready to load into Amigo. Approximately half the genome, everything other than hypothetical protein-encoding, has been annotated (2400 genes). It includes ~500 annotations to the malaria specific component term "apicoplast". Many of these annotations have now been experimentally confirmed and given an IDA evidence code. GeneDB now has funding for a malaria curator.

Over the next few months GO annotation of the genomes of Aspergillus fumigatus and Theileria annulata will commence. Joint curation with TIGR of Trypanosoma brucei will continue.

More Tsetse EST sequencing is also planned.

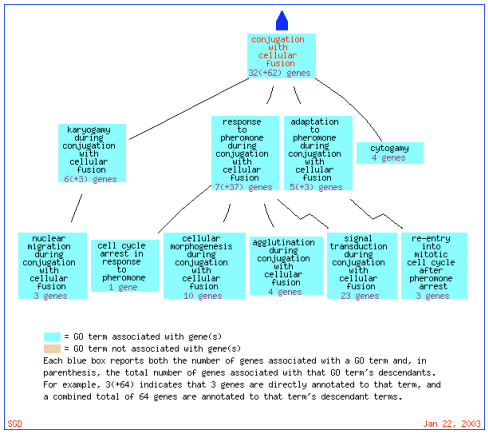
SGD Progress Report January, 2003

General goals

- Increase awareness of GO among yeast community
- Create tools that allow users to take advantage of the GO annotations
- Complete annotations of all *S. cerevisiae* gene products
- Continued development of the ontologies, including definitions

Recent developments

- Every ORF at SGD has been annotated to Function and Process terms
- Every named ORF has a complete set of GO annotations
- Released **GO Term Finder**
  - ** finds the most significant GO term that a list of yeast genes shares in common
  - ** http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermFinder
- Released **GO Tree View**
  - ** visual representation of the DAG
  - ** used for Gene Ontology term page
  - ** used for GO Term Finder
  - ** can be used to browse ontology
  - ** shows # of genes annotated to term
  - ** shows # of genes annotated to children
- GO ontologies loaded into database every night



Items in progress

- Have complete GO annotations for all SGD ORFs
  - ** Component terms for 786 unnamed ORFs
  - ** Review annotations that have IEA as an evidence code (<40 for named ORFs)
- Preliminary stages of creating a GO-Slim with yeast specific processes
- Continued efforts in defining terms
  - ** each curator defines 2 terms/month
  - ** giving priority to terms that have been used to annotated genes at SGD
- Developing ontologies
  - ** during the course of the SGD GO jamboree and during the creation of the yeast-specific GO-Slim, found areas that need development and expansion
- Will release the yeast metabolic pathway using Peter Karp's Pathway Tools
  - ** used the EC2GO mapping
- Creating an advanced search that allows users to find genes that are annotated to the intersection of 2+ GO-Slim terms

Appendix 1H. TAIR

TAIR progress report

January, 2003

IEA Annotation at TAIR:

InterPro2GO:

InterProScan.pl was run on all Arabidopsis proteins (ATH1.pep, version July 2002), using the Interpro database version 4.0.

TargetP prediction of subcellular localization:

TargetP was run on all Arabidopsis proteins (ATH1.pep, version July 2002). This resulted in about 11,500 component annotations.

Metacyc2go:

A mapping file, metacyc2go, was used to generate the GO annotations. This resulted in about 1800 annotations

NON-IEA annotation at TAIR:

* Number of unique genes annotated: 3175
* Number of annotations: 5134

***New strategy**: **TAIR terms** in GO (490)->Arabidopsis genes with valid article hit.
* Similar strategy, but to **non-TAIR terms** in GO->Arabidopsis genes with valid article hit
*Currently using anatomy and developmental stage ontologies to annotate expression pattern.

'Go annotation':	TAIR-term	Non-TAIR-term
Number of genes (October)	900	1500
Number of genes (January)	485	-----
'Anatomy and Developmental stage':		
Number of genes	112	

* Goal for 2003: Annotate all studied Arabidopsis genes (the ones with the literature associations) to all three GO ontologies.

Non-annotation issues from TAIR:

*GO terms added to process and function ontology with definitions.
* Added Plant GO slim to Goslim directory at ftp://ftp.geneontology.org/pub/go/GO_slims/
* Updated TAIR's gp2protein file on GO
* Gene-association file includes PMID/Agricola as secondary id
* Cellular process terms->work in progress by Tanya and David

Appendix 1I. TIGR

TIGR Eukaryotic GO update January 24, 2003 Linda Hannick

The *Arabidopsis thaliana* project has temporarily halted assignment of GO pending completion of some required tasks for our grant. Work on GO will resume when these tasks are complete. GO ID's will be assigned to *Trypanosoma brucei* chromosomes 4 and 6, beginning in the next month or so. Other projects in the planning stages include rice (*Oryza Sativa*) and *Aspergillus fumigatus*.

TIGR Prokaryotic GO update January 24, 2003March 3, 2003 Michelle Gwinn

We have just sent to GO the association file for *Shewanella oneidensis*. This contributes 8292 GO term associations to 3767 prokaryotic genes. We still have several bacterial genomes with GO associations awaiting publication and subsequent release of the GO data that will add greater than 20,000 associations to greater than 10000 genes. These include *Bacillus anthracis*, *Listeria monocytogenes*, and *Methylococcus capsulatus*."

Action Items from Cambridge Sept 2002 meeting

1. FB to use PubMed IDs instead of [or in addition to?] FBrf IDs.
- DONE

2. TIGR to provide protein id --> TIGR gene ID.
- DONE

3. TIGR to send IEA annotations to GO for genomes not sequenced at TIGR.
- NOT DONE. Michelle says some of IEA associations were being made based incorrect GO associations and is working to fix this.

4. Cath will update documentation and circulate drafts.
- see report

5. Evelyn to continue tracking down info on QuickGO concurrent assignments.
- has tried. contact David Binns.

6. Consortium, especially Chris M, to revisit concurrent annotations in GO database.
?????

7. Add check for term deletion to flat file helper.
- will put option in configuration manager

8. Sue will ask Danny to take over DAG-Edit maintenance.
- DONE. He said no.

9. Amelia will collect bug reports and feature requests for DAG-Edit from curators. If John can't act on feature suggestions, perhaps Danny can.
- DONE. SourceForge list.

10. Change prefixes to "GOC:" for definition references that represent an individual curator or group of curators.
- when Brad does 11.

11. Brad will create a form where curators can enter info (e.g. name, affiliation, dbxref entered in definition reference field), and create and link a web page for each GOC:xyz entry.
- new action item covering this.

12. Chris to get comments into the database.
- code working. will do.

13. Add a link to the GO-Slim directory to the home page.
- NOT DONE.

14. DBs to send GO-Slims and lists of all genes to BDGP.
- in directory.

15. BDGP to generate tables of gene ID <--> GO-Slim term for each DB that submits a gene list and a GO-Slim. Genes lacking annotations will get "unexamined"; annotations to "unknown" will be preserved.
?????

16. Add hyperlinks to the gp2protein files: link from web page and from each gene_association file.
- use docs

17. Set up "interest groups" based on subject matter; maintain a list of groups and who's in them (on SourceForge if possible -- look into this).
- sort of DONE.

18. All content changes, no matter how small, should go into the SourceForge tracker for archiving purposes. Summary entries should be nice and informative.
- ongoing; DONE.

19. Set up script to email summaries from new (open) SourceForge tracker entries.
- DONE.

20. Test all "protein biosynthesis" and "protein binding" terms. Apply the two-part test to all, and (for protein family or class ones) look at annotations and child terms. Circulate the list slated for obsolescence. Note: we are not going to make all "protein binding" terms obsolete yet. It would be good to determine which terms would pass the tests, though.
- in progress.

21. Circulate a proposal for incorporating "gene expression" and "regulation of gene expression" terms and definitions.
- decided against "regulation of gene expression"; Jane will circulate the "gene expression" def.

22. Discuss this [protein binding etc.] again at the next meeting!
- DONE.

23. Propose definition for "cellular process" and discuss on mailing list.
- DONE.

24. Each model organism DB should review terms under "embryogenesis" and "morphogenesis" to check for correct parentage; also figure out which ones will go under "cellular process."
- in progress; mouse done.

25. TAIR curators to improve definitions of "cell surface" and its children.
- DONE.

26. Change wording of GO:0030312 to "external encapsulating structure." Circulate new definition; make sure Michelle Gwinn has a chance to comment.
- DONE.

27. Review all "cell wall" terms to check parentage. Plant cell wall does need to be moved.
- DONE.

28. Start thinking about terms (and definitions, of course) to capture concept of boundary.
- ongoing.

29. Create UniGene <--> GO file (Daniel)
- DONE.

30. Add to documentation of "with" column use -- allow cardinality 0, 1, >1 for all evidence codes that use "with" at all; explain situations where cardinality 0 is allowed.
- NOT DONE.

31. Annotations that use ISS, IPI, or IGI but have a blank "with" column should link to the annotation documentation (let people see the possible reasons why nothing's entered).
- NOT DONE.

32. Each group that shares annotations should tag the ones that come from the other group(s).
- coming soon.

33. Document this decision [shared annotation], and how to implement it.
- coming soon.

34. Amelia will continue polishing The Script. When it's ready for prime time, it will go in the software repository, and will be run every month to generate a log to accompany the flat file archives and database releases. Decide where to put the output.
- script done; need to decide where output should go.

35. set up new faq-o-matic page (Cath & Rama, with a bit of help from Chris); everyone to add faq's and answers, though Cath & Rama will probably do the most, at least at first.
- content collection 1st round done.

36. EBI GO curators circulate a set of instructions for using CVS.
- DONE.

37. Progress report for current grant.
- DONE.

38. Prepare renewal grant application.
- in progress.

39. Prepare a site with mock-ups of GO web pages derived by splitting up the current home page sensibly.
- NOT DONE.

Integrating GO into UMLS – the story so far…

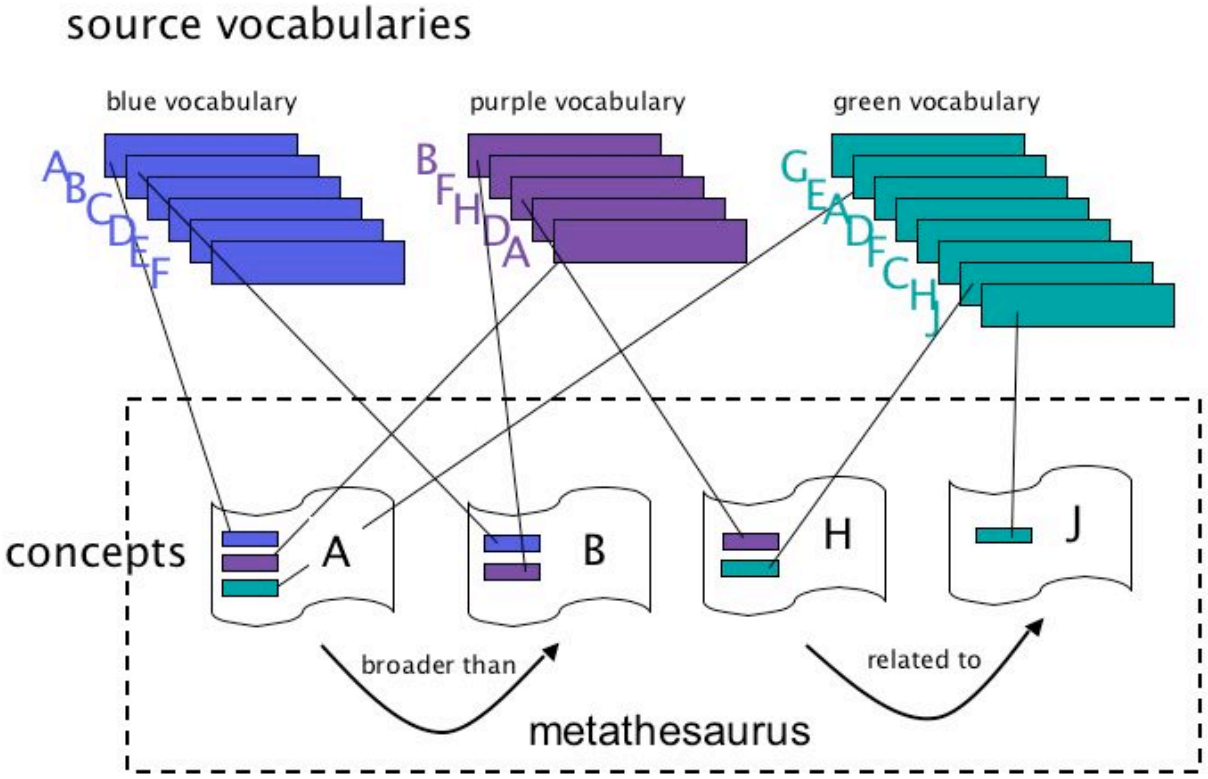
- Overview:
  - First insertion September 2002 – all ontologies
  - My visit to NLM Sept/Oct 2002
    - Few problems with insertion
    - Demoted insertion – more fiddling
  - Second insertion January 2003 – molecular function only

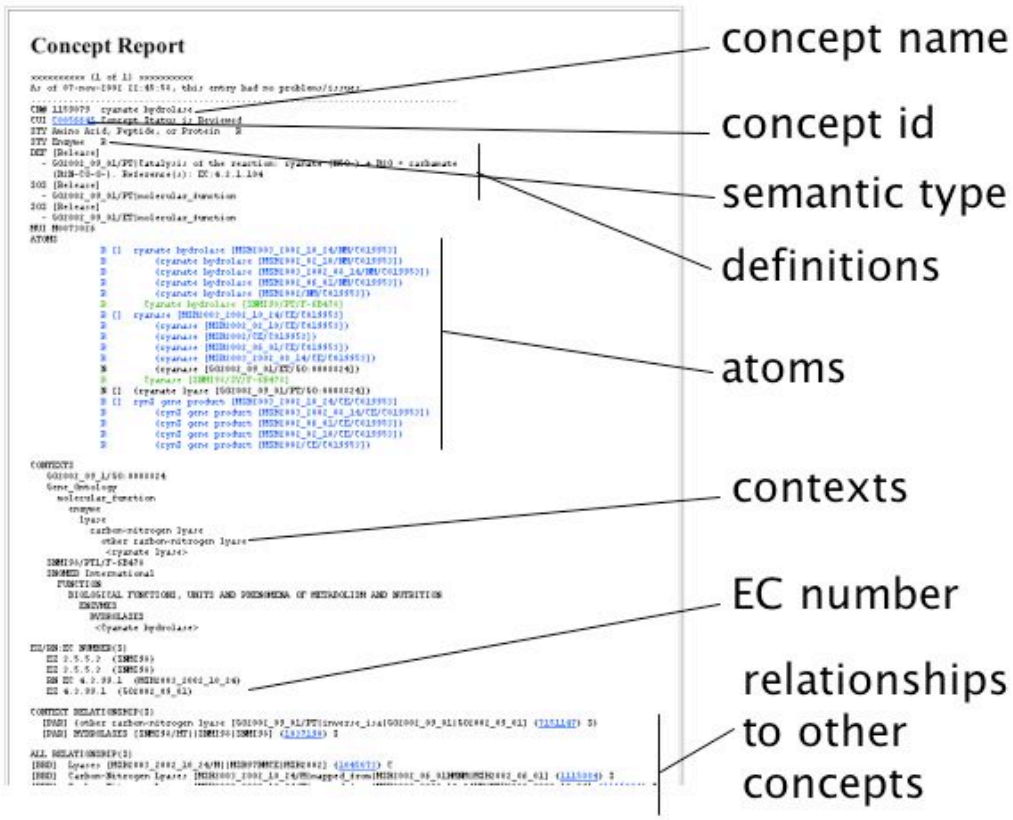
Unified Medical Language System (UMLS)

- Research project maintained by the National Library of Medicine (NLM)
- Has three parts – 'Knowledge sources':
  - UMLS Metathesaurus
  - SPECIALIST lexicon
  - semantic network

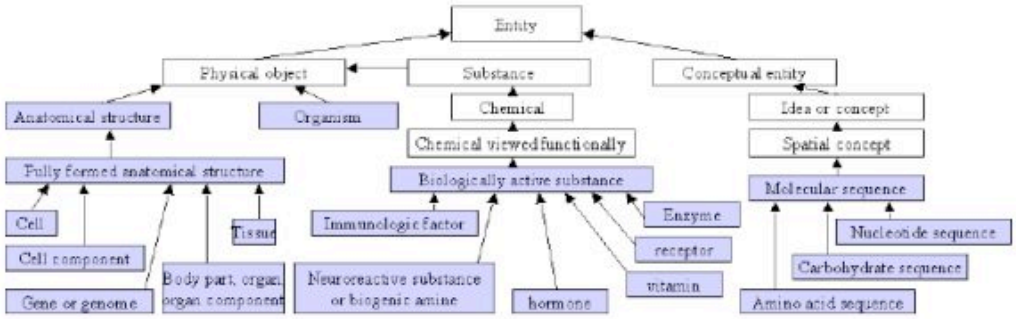
UMLS Metathesaurus

- Resource for making links between different controlled vocabularies.
- 'Terms' (concepts) in metathesaurus can have relationships to one another
- Vocabularies include MeSH, SNOMED, CRiSP thesaurus.





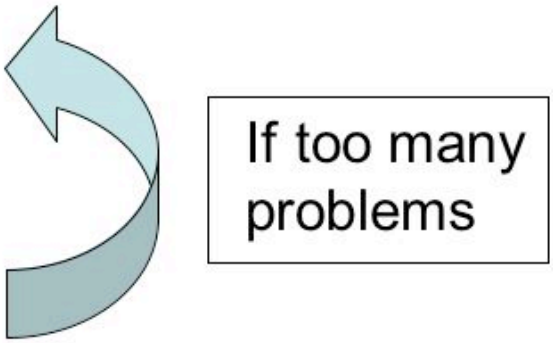
The semantic network



- physically related to
- spatially related to
- temporally related to
- functionally related to
- conceptually related to

Inserting a new vocabulary into UMLS

- Convert to correct format (inversion)
- Make rules/defaults
- Insertion
- Hand editing



Inserting GO into UMLS (take I)

- Limited matches by semantic type
  - e.g. never match a GO term to a concept with the semantic type 'Medical device'
- Used synonyms and dbxrefs to aid matching
  - Called synonyms exact

Problems (take I)

- Synonyms not synonymous
  - Caused GO terms to be inserted into inappropriate concepts
  - Same synonym for multiple terms = horrible mess – different terms matched together
  - Demotions
- GO functions/component v/s protein entities
  - Merging GO function/component terms into enzyme concepts didn't work – semantic type aa, peptide or protein
  - Many other function terms same problem – molecules
    - receptors
    - carriers
    - structural molecules

Solutions – synonymy

- Don't use GO synonyms in UMLS
  - GO synonyms very useful in matching process – reduce false negatives
- Okay, so classify GO synonyms so we can differentiate the exact ones
  - Good idea! Will probably take a while though.
  - Alright, well create a version of GO with all non-exact synonyms removed

Solutions – MF terms

- Make it illegal for GO MF terms to match concept which is an entity
  - Leads to two concepts with the same name – not good in any vocabulary
- So, change the name of the GO term
  - Okay, appended 'activity' to all GO MF terms (except binding and a few others) in version of GO used for match
  - Problem is that now, terms in UMLS different to those in GO
  - Change in GO?

Insertion of GO into UMLS (take II)

- Molecular function only
  - added activity
  - removed all non-exact synonyms
  - restricted matches to concepts with following semantic types only:

Natural Phenomenon or Process
Biologic Function
Physiologic Function
Organism Function
Mental Process
Organ or Tissue Function
Cell Function
Molecular Function
Genetic Function

Problems (take II)

- Will need to produce a non-exact synonyms file for each UMLS release
  - Hold information in GO somehow
    - Database only?
    - If flat files, how will this affect tools?
- Different text strings in GO and UMLS
  - Add 'activity' to GO MF terms, retain original as synonym?
  - Add 'activity' version of term as a synonym?

What next?

- Edit and (hopefully) release GO MF with UMLS in March
- Insert, edit and release BP and CC (June/July)
- Resolve synonym issue in GO – adjust insertion recipe
- New insertion of GO for each UMLS release (quarterly) – same recipe

Appendix 4A. Email from Tanya Berardini

Cellular process issues (from Tanya):
Subject: Cellular process issues for St.Croix

Hi everyone,

Here are a few issues that I think would be good to address at the meeting. David will be attending, while I won't be able to make it.

1. cell differentiation vs. cell fate commitment
right now, these terms are siblings
cell differentiation: The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.
ref:ISBN:0198506732

cell fate commitment: The commitment of cells to specific cell fates and
their capacity to differentiate into particular kinds of cells. Positional information is established through protein signals that emanate from a localized source within a cell (the initial one-cell zygote) or within a developmental field.
ref: ISBN:0716731185

2. response to endogenous stimulus and response to exogenous stimulus

Move to be children of physiological process/add physiological process as additional parent? Right now, they are children of cell communication.

response to endogenous stimulus: The change in state or activity of a cell or an organism as a result of the perception of an endogenous stimulus.
ref: TAIR:sm

response to exogenous stimulus:The change in state of activity of an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of the perception of an external stimulus.
ref: FB:hb

3. cell_type development vs. cell_type differentiation
Do we need both terms? Are they meant to describe different things? (e.g. pole cell development vs. pole cell differentiation) Check out the children of cell differentiation for a sample.

Thanks,

Tanya

Appendix 4B. Email from Aubrey De Grey

Term grammar (from Aubrey de Grey):
Subject: GO grammar

Hi Midori,

Am I alone in feeling that the GO ontologies are grammatically challenged? It seems to me that the terms in each of them should be such that a sentence of the form:

It encodes a[n] <function> involved in <process> which is localised
to the <component>

should always read properly, but in fact one gets things like:
It encodes a heme binding involved in nutritional response pathway
which is a component of the extracellular.

as opposed to:
It encodes a heme binding protein involved in nutritional response
which is a component of the extracellular space.

I care about this more than most because I construct such sentences automatically from GO data in FlyBase as part of the summary paragraphs that appear in the gene records. But I think it looks decidedly untidy even when the terms are presented in tabular form, and it would probably take only a couple of hours' work to correct the common ones. Becky saw
my point and suggested I mention it to you. What do you think?

Cheers, Aubrey

reply:
Hi Aubrey,

I'll put this issue on the agenda for the GO meeting, since it's coming up so soon anyway. I don't think there' will be any objection to adjusting the 'pathway' process terms, or the cellular component terms, since it will help with sentence generation, and won't hurt for any other purpose.

We absolutely _cannot_ make alterations such as 'heme binding' --> 'heme binding protein' in the function ontology. None of the ontologies includes terms representing gene products; rather, we did and do put a lot of effort into keeping gene product names (whether specific, like 'actin', or generic, like 'protein') out of GO. GO terms also do not represent what a
gene product is (or is made of), but what it does and where it is found. Function terms represent activities, not entities.

It seems to me that it would be straightforward to adjust the sentence generation to accommodate function terms as activities rather than molecules, e.g.

It encodes [an RNAla protein] with <function> activity involved in ...

We would then be willing to fix any function terms that caused this construction to go awry.

reply to above:

Very good point re function - and very nice suggestion for the sentence structure. That's what I'll do. On a quick browse, the only group of function terms that would be a bit broken by your sentence structure are ones that end in "factor" (guanyl-nucleotide exchange factor, etc), and for them I guess dropping "factor" would actually be in line with the policy you describe. Great if you can adjust the process and component ontologies.

Appendix 4C. MGI Excessive granularity document

Excessive granularity.

As we add and refine terms in the ontologies, we need to keep two things in mind. First, the terms should be as organism non-specific as possible. Secondly, the terms should be as meaningful as possible. As put forth in the last GO meeting, there are several branches of the GO that seem to have expanded unnecessarily. These fall into three broad categories: Protein Binding, Biosynthesis, and Regulation.

1. Protein binding:

In using the term GO:0005515, Protein binding, we can make use of not only the GO ontology structure itself, but also the use of the attributes/qualifiers used in linking a term in the ontologies with a gene product, which are included in each annotation line supplied in a gene_association.db file.

A good example is the use of the term "protein binding". This term can be qualified with both an evidence code and the "with" field. The combination allows a curation of a gene product to bind to a specific protein product. The "with" field is intended to house a sequence identifier or db identifier pointing to a specific protein. **Therefore, there should be no need to populate the GO with specific children of protein binding**. However, that is not to say that in those instances where there may be ambiquity, that we cannot have a child that describes binding to a product family. For example, actin binding, stat binding, etc. These can be used when the specific gene product is not identified.

Amel, amelogenin
F GO:0005515 protein binding IPI SWP:Q9CRG8
In this example, amelogenin was shown to bind to Q9CRG8, the protein specified by Bat3, HLA-B-associated transcript 3

Acrp30 adipocyte complement related protein of 30 kDa
F GO:0005515 protein binding IPI SWP:Q60994

In the example above, the protein Acrp30 is shown to bind to SWP:Q60994, Acrp30; thus, the statement demonstrates that the protein oligomerizes.

Ablim1, actin-binding LIM protein
F GO:0003779 actin binding IDA
In this example, the actin-binding LIM protein was shown to bind actin, but the actual gene product was not specified (in mouse, there several actins: actin, alpha 1 (Acta1), actin alpha 2 (Acta2), actin beta (Actb), actin alpha, cardiac (Actc1), actin gamma (Actg), and actin gamma2 (Actg2). The term GO:0005515 would not be sufficient, since the "with" field could not be specified. However, in this case the GO:0003779 term allowed sufficient granularity in the annotation.

Another example uses GO:0005518, collagen binding.

Gp6, glycoprotein 6 (platelet)
F collagen binding IDA
In this example, glycoprotein 6 was shown to bind (a) collagen.

However, in the case of Mrc2, mannose receptor, C type 2

F collagen binding ISS EMBL:AF107292
a human ortholog of the murine mannose receptor was shown to bind collagen. In this instance, it it NOT the mouse protein that was assayed, so it would be inappropriate to use the human binding target. However, we infer that because the paper shows that AF107292 is the human ortholog of the mouse protein, we can assign the collagen binding function. Again, because a suitable child existed for the protein binding term, we can capture the protein binding function with more granularity than would otherwise be possible.

Therefore, in most cases, the use of the "with" field in combination with the IPI code is sufficient to annotate binding of one protein to another. It is therefore not necessary to consider creating protein-specific terms (eg, interleukin 1-15 binding) to capture the information.

2. Biosynthesis

As maintained before, the notion of Protein Biosynthesis should mean specifically the building up of a polypeptide by translation. Any other fate of the protein, such as post-translational modification, etc. is NOT part of "Protein Biosynthesis". The use of the term Biosynthesis to include other metabolic fates is misleading. Protein biosynthesis is already itself a child of metabolism.

Thus, adding terms such as "biosynthesis of protein X" as a term to mean anything affecting the appearance/level of protein X is not useful. If a gene product effects the translation of protein X, then the gene product's annotation should be to a specific term under protein biosynthesis (initiation, elongation, etc.). If the gene product effects/ modifies a post-translational modification, etc., then it should be annotated to those processes.

3. Regulation

Additionally, terms are arising in several notes concerning the regulation, both positive and negative, or particular processes (biosynthesis, phosphorylation, etc.). Terms exist for the negative/positive regulation of phosphorylation/whatever of specific_protein_family_member X, X+1, etc. Is this granularity necessary? Would it be sufficient for negative/positive reglation of phopsphorylation/whatever period/or protein_family?

Protein Biosynthesis Example:1

protein biosynthesis [GO:0006412]
- amino acid activation +
- charged-tRNA modification +
- **glycoprotein biosynthesis+
  - CD4 biosynthesis +
  - FasL biosynthesis +
  - protein amino acid glycosylation +
- *integrin biosynthesis + and children
- **lipoprotein biosynthesis and children+
- **mannoprotein biosynthesis and children +
- *MHC class I biosynthesis and children+
- *MHC class II biosynthesis and children+
- *neurotransmitter receptor biosynthesis
- non-ribosomal peptide biosynthesis
- regulation of protein biosynthesis +
- regulation of translation +
- **TRAIL receptor biosynthesis and children**
- translational elongation +
- translational initiation +
- translational termination +

viral protein biosynthesis

Biosynthesis Example 2

immune response
cytokine metabolism
cytokine biosynthesis
chemokine biosynthesis +
connective tissue growth factor biosynthesis +
granulocyte macrophage colony-stimulating factor biosynthesis +
~~interferon type I biosynthesis +~~
~~interferon-gamma biosynthesis +~~
~~interleukin-1 biosynthesis [GO:0042222]~~
~~regulation of interleukin-1 biosynthesis +~~
~~interleukin-10 biosynthesis +~~
~~interleukin-11 biosynthesis +~~
~~interleukin-12 biosynthesis +~~
~~interleukin-13 biosynthesis +~~
~~interleukin-14 biosynthesis +~~
~~interleukin-15 biosynthesis +~~
~~interleukin-16 biosynthesis +~~
~~interleukin-17 biosynthesis +~~
~~interleukin-18 biosynthesis +~~
~~interleukin-19 biosynthesis +~~
~~interleukin-2 biosynthesis +~~
~~Interleukin-20 biosynthesis +~~
~~interleukin-21 biosynthesis +~~
~~interleukin-22 biosynthesis +~~
~~interleukin-23 biosynthesis +~~
~~interleukin-24 biosynthesis +~~
~~interleukin-25 biosynthesis +~~
~~interleukin-26 biosynthesis +~~
~~interleukin-27 biosynthesis +~~
~~interleukin-3 biosynthesis +~~
~~interleukin-4 biosynthesis +~~
~~interleukin-5 biosynthesis +~~
~~interleukin-6 biosynthesis +~~
~~interleukin-7 biosynthesis +~~
~~interleukin-8 biosynthesis +~~
~~interleukin-9 biosynthesis~~ +
regulation of cytokine biosynthesis +
~~TRAIL biosynthesis~~ +

Regulation Example

regulation of tyrosine phosphorylation of STAT protein
positive regulation of tyrosine phosphorylation of STAT protein
~~positive regulation of tyrosine phosphorylation of Stat1 protein~~
~~positive regulation of tyrosine phosphorylation of Stat2 protein~~
~~positive regulation of tyrosine phosphorylation of Stat3 protein~~
~~positive regulation of tyrosine phosphorylation of Stat4 protein~~
~~positive regulation of tyrosine phosphorylation of Stat5 protein~~
~~positive regulation of tyrosine phosphorylation of Stat6 protein]~~
~~positive regulation of tyrosine phosphorylation of Stat7 protein~~

NOT Protein Binding

Background:

The GO term "protein binding" (GO:0005515) is used in the function ontology to specify that a gene product binds to another protein. It is used with the IPI evidence code and the "with" field to indicate the specific protein that the annotated gene product binds to.

In the examples below, Arl6ip has been shown to bind to Arl6 (SP:O88848), and Cdc42 has been shown to bind Cdc42ep5

Arl6ip ADP-ribosylation-like factor 6 interacting protein

```
F     protein binding    IPI    SP:O88848

Cdc42,       cell division cycle 42 homolog (S. cerevisiae)
F     protein binding    IPI    SWP:Q9QZT9
```

The "not" qualifier has been provided for documentation of experiments that were designed to test a hypothesized function, cellular localization, and proposed participation in a biological process. For example, a protein product has homology to chitinase; however, experiments performed on the isolated protein demonstrated that the protein did NOT have chitinase activity.

```
Chi3l3       chitinase 3-like 3-
F     NOT     chitinase   IDA
```

Dilemma:

In some experiments, protein binding to a specific protein has been shown to **not** occur. In the example below, a publication demonstrated that the gene product of Akap9 specifically binds one protein, but not the other.

```
Akap9        A kinase (PRKA) anchor protein (yotiao) 9
C            cytoplasm    IDA
F     NOT    protein binding    IPI    SWP:Q62348
F            protein binding    IPI    SWP:Q9QZE7
```

However, in this instance, the use of the "NOT" may be confusing, as the GO term "protein binding" is (probably) meant to be very broad (it has no definition), and does (may) not imply "binding to a specific protein". For example, immunoprecipitation experiments could demonstrate that a particular gene product is associating with other proteins, but the proteins have not been identified. In this case, the "with" field may have to be left null. **The risk, however, is that the "not" could be misinterpreted to mean that this gene product does NOT have the function of binding to a protein.**
Generally, when an assertion can use the "with" field, the annotation still makes sense if that field is blank. For example, when an ISS evidence code is used, but the accession number is not known, leaving the "with" field blank still means that the annotation was made based on sequence similarity. Another example is when the IMP evidence code is used. If the assertion is based on a specific mutant allele, it is possible to add a database identifier to the "with" field, when known. However, if the assertion is based on an RNAi experiment, the "with" field is often left blank. In these cases, the annotation makes sense even if the "with field is blank.

A problem can arise, however, if the "not" qualifier is used with protein binding and IPI. If the "with" field is left blank, the assertion reads that the gene product does not bind protein. Note that this is not a problem when a gene product can be annotated to one of the children, such as "actin-binding" (does NOT bind actin).

Proposal

We would still like to be able to capture this type of experiment, as it can provide information about the properties of the gene product. Therefore, it might be useful to create a term , such as "specific protein binding" as a child of protein binding. All/most of the current children of "protein binding" would then be moved to be children of the new term. The "not" qualifier would never be used if the "with" field is left blank.

For example, the entry for Kdr is shown below:

Kdr, kinase insert domain protein receptor
F NOT specific protein binding IPI SP:P97946

The interpretation should be that Arl6ip does not bind specifically to Figf (c-fos induced growth factor).

A second example, is where

Akap9, A kinase (PRKA) anchor protein
F NOT specific protein binding IPI SWP:Q62348
F specific protein binding IPI SWP:Q9QZE7

This paper demonstrated that Akap9 did NOT bind to Tsn (translin), but DID bind to Tsnax (translin-associated factor X).

Appendix 4E. Documentation progress report (from Cath)

GO DOCUMENTATION: PROGRESS REPORT

OVERARCHING PRINCIPLES

- Make as much as poss. comprehensible to broad audience
- Make it clear what audience each doc is aimed at
- Avoid redundancy wherever possible to make pages easier to update (FAQ is an exception to this principle: aim is to provide info with as few clicks as possible)

STUFF FOR THE GENERAL PUBLIC

An introduction to GO

- Purpose is to provide an overview that is clear and useful to first-time users. Links to more detailed documents make it useful for curators and annotators.
- Includes General documentation up to Data representation section: defines what's covered (and what isn't) in each ontology, plus the basics of DAG structure (will redo diagram so it looks better on the web)
- Replaced 'data representation' with a blurb about what file formats we produce and where you can download them from. Also includes a para on GO slims.
- New section that puts GO in the context of ontologies in general: discusses GOBO and the new list of ontologies on the MGED site; dicusses cross products, and discusses mappings to other classification systems.
- 'Contributing to GO points to the sourceforge site and to the mailing lists.
- Still needs a link to the FAQs.

FAQs

- Have kept html v. simple and not added comprehensive contents list yet because these will be pasted into FAQomatic and this might do some of the formatting for us.
- Where are the main gaps and who can provide material to fill them?
- What other sections do we need and what order should the sections be in?
- Do we need a section on annotations to each of the MODs or should these be dealt with by the MODs' own FAQs? If so, GOA FAQs could be moved to GOA web page and we could just provide links to each MOD's FAQ.
- Should the order of questions in any of the sections be swapped around?
- Need to install the faqomatic (http://faqomatic.sourceforge.net/fom-serve/cache/1.html) . Can Chris do this?
- Do we need someone to be in charge of the FAQ or should it be a free for all?
- Can each of the people who provide questions check them: in some cases I've made them more general and added bits.

GO style guide (or should it be the GO content guide?)

- Revamp of GO usage guide. Purpose is to explain not only how the ontologies are created and edited, but also the rationale behind why we do it this way.
- I wrote it as a practical guide for curators but I don't think it's working.
- Starts off at the level of terms and what we can do with them; then moves up a level to relationships between terms; then deals with whole ontologies and the rules specific to each one.
- Should we split the purely philosophical (more in tune with the original purpose of this doc) from the purely practical? If we did this, people who aren't part of the consortium but want to know more about *why* we do it that way would have something deeper than the intro.
- If we did this, should we merge the purely practical stuff with the format guide? This would avoid the constant cross-referencing between docs.
- Things that have been added include

  Much more comprehensive contents list

link to full list of database cross-references
Amelia's list of 'standard' definitions
Clearer guidelines on sensu

GO format guide

- Main aim to help anyone who wants to parse the files; but also of use to curators because you always end up tweaking the flat files at some point.
- As with the style guide, it starts of at the level of terms, then moves up to relationships between terms and the structure of entire files.
- Not sure what to do with the stuff on the bibliography.
- Need someone to write up something on structure of mySQL files, or provide a link if this is already there on the godatabase site.
- Added Jane's syntax for comments
  More on how to use sensu.

Stuff that I haven't done….(mention at this point that I'm now full time outreach)

Publications on/about GO

- Update; reverse order so most recent first.

STUFF FOR CURATORS

CVS user guide for curators

- new doc; I'm happy to edit this but I need someone to write it. Volunteers?

DAG-Edit User Guide (Jane's doc)

- Does this need to be more visible from the front page now that other ontologies are using it more and more?
- Jane to add some info on creating cross products
- Needs formatting in same style as rest of docs; I'm not going to do anything else with it.

Dummies' guides

- Each group to maintain their own local 'dummies' guides': SGD and EBI now have these.
- I'll turn the EBI one into HTML and leave it on my website (or the website of one of the other curators? I'll have to pass over the responsibility of updating this to someone else.

STUFF FOR ANNOTATORS
GO Annotation Guide

- Computational annotation methods need updating.
  FlyBase (Becky Foulger)
  SGD (Karen Christie)
  MGI (Harold Drabkin - has already sent info to Midori)
  TAIR (Suparna has sent)
  WormBase ?
  PomBase (Val Wood)
  RGD ?
  DictyBase (Rex Chisholm)
  PSU (Matt Berriman)
  Gramene (Pankaj Jaiswal)
  GKB ?

EBI (Daniel Barrell)
TIGR (Michelle Gwinn/Linda Hannick)
Compugen (liat Mitz/Han Xie)
AstraZeneca Courtland Yockey
Incyte Lisa Matthews

- I'll add the ones I've got but then I'd like to hand over to someone else.
- Need to document standard operating procedures for shared annotations (tag annotations that come from other groups).

Appendix 5. Collected action items from this meeting

Action Items, St. Croix January 2003

1. TAIR. Update MetaCyc2GO mappings.

2. John. Action item 7 from last time [add term deletion feature to DAG-Edit].

3. Brad. Action items 10 and 11 [adding more information about GO curators to website/database] outstanding.

4. Come up with system for notifying developers of format changes.

5. Add "contributed by" column.

6. Curators. When adding new synonyms, track which type they are. If they are ‘broader than’ or ‘narrower than’, consider whether it calls for a new term.

7. Jane. Circulate synonym list again.

8. BDGP. Look into rules that could be worked into DAG-Edit to make synonym maintenance easier.

9. Jane. Discuss this with UMLS and fill us in on the results.

10. David and Tanya. When splitting out multicellular v/s unicellular processes, make the split as far below ‘physiological process ; GO:0007582’ as possible, and as and when needed, rather than splitting right below physiological processes.

11. GO editorial team (and others). Start removing grouping terms slowly and carefully with all the usual communications. If obsoleting a term, ensure the corresponding process or component exists.

12. Jane. Add activity to function term strings.

13. GO editorial team. Define extracellular to include outside a virus particle, then use host terms as parents for the appropriate virus cell component terms.

14. GO editorial team. Go through the enzyme complexes (see also SF entry 535294) and where applicable, make a general parent directly under ‘cellular component’ with children in specific locations.

15. GO editorial team. Get a list from Aubrey of ill-fitting GO terms and evaluate; adjust terms as needed.

16. Announce on the website that we'll implement this solution at some future date (no date set but will be 6+ months from now). Assemble a group (MA, Chris, David) to work on the implementation.

17. Midori. Put up interest groups on web page. Everybody to send group ideas & which they volunteer for. See if it works or if we need further formalization by putting groups in SourceForge.

18. Construct and post a user survey covering tools, AmiGO, etc.. Send question ideas to Amelia Ireland. It will be sent out to GO-Friends and data collected in time for the grant application.

19. Chris. Suggestion: a daily release of a separate database containing just terms without annotations. The whole database should be updated every month. AmiGO would have the option to view the up-to-date term set with no associations.

20. Chris. Make use of parents rather than bucket terms to avoid confusion due to transient IDs.

21. Brad. Investigate piping GO-Slim mapping results to the AmiGO pie chart maker.

22. Brad. Add the ability to dump AmiGO pie chart data as a flat file containing GO ID, term name and the number of gene products.

23. Member databases. Each database should send annotation FAQs from their existing documentation to Cath for inclusion in GO FAQ. GO FAQ will have general annotation FAQs and then specific FAQs from each database and from the EBI.

24. Everyone . Read over the new documentation (especially the style guide) and send any suggestions to Cath. This is available at http://www.ebi.ac.uk/~cath/ .

25. Cath. The changeover to the new documentation will occur on 15 March.

26. Cath. Update the synonym section of format guide to accommodate the decisions made at this meeting.

27. Chris. Provide some documentation on the mySQL database.