Gene Ontology Consortium Meeting

Lucy Cavendish College, Cambridge, UK
September 10-11, 2002

Contents

Participants:

Michael Ashburner	FlyBase	Cambridge, UK
Rama Balakrishnan	SGD	Stanford, CA
Daniel Barrell	EBI	Hinxton, UK
Tanya Berardini	TAIR	Carnegie Inst., Stanford, CA
Matt Berriman	PSU(Sanger)	Hinxton, UK
Judith Blake	MGI	Bar Harbor, ME
Cath Brooksbank	EBI	Hinxton, UK
Evelyn Camon	EBI	Hinxton, UK
Mike Cherry	SGD	Stanford, CA
Rex Chisholm	DictyBase	Northwestern Univ., Chicago, IL
Karen Christie	SGD	Stanford, CA
Bernard de Bono	MRC-LMB	Cambridge, UK
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
Suzanna Lewis	BDGP	Berkeley, CA
Jane Lomax	EBI	Hinxton, UK
Brad Marshall	BDGP	Berkeley, CA
Lisa Matthews	Incyte Genomics	Beverly, MA
Suparna Mundodi	TAIR	Carnegie Inst., Stanford, CA
Chris Mungall	BDGP	Berkeley, CA
Sue Rhee	TAIR	Carnegie Inst., Stanford, CA
John Richter	BDGP	Berkeley, CA
Erich Schwarz	WB	Caltech, CA
Valerie Wood	PomBase(Sanger)	Hinxton, UK
Han Xie	Compugen	Jamesbrook, NJ

Visiting on Tuesday, Sept. 10, 2002:

Robert Stevens	University of Manchester
Chris Wroe	University of Manchester

Progress Reports

GO Curators at EBI

- Jane to visit NLM (more below)
- 60% of terms now defined
- MIPS Funcat <—> GO mapping posted (go/external2go/mips2go)
- other aspects of progress touched on in review of action items from CSH

FlyBase

- see handout; highlights:
  - recuration for release 3 of *Drosophila* sequence (gaps filled; new genes)
  - Eleanor Whitfield (SP) cross-checks FB & SP annotations for redundancy

** **action item 1**: FB to use PubMed IDs instead of [or in addition to?] FBrf IDs

SGD

- see handout; highlights:
  - new GO Term Mapper and GO Term Finder tools
  - GO Tutorial (some parts generic GO, others SGD-specific)
  - at least one annotation for every gene known to encode a product

MGI

- see handout; highlights:
  - areas where GO annotation is focused
  - cross-product manuscript accepted (Genome Research)
  - work on cellular and developmental processes

TAIR

- replacing IEA with literature-based annotations
- nifty cell viewer
- organize annotation effort by cellular component (using cell viewer) or by pathway
- Pubsearch tool helps with literature mining (from Suparna's Users Meeting talk)

WormBase

- developmental stage ontology to be released soon (waiting for some data on aging to be made public)
- anatomy ontology also in the works; has about 5900 terms!
- working on tool, way to handle GO annotations in ACeDB
- will update RNAi —> GO term mapping (used for some WB IEA annotations)

DictyBase

- NIH funding started August 1
- SGD tables loaded with Dicty data
- manual curation getting started

PSU

- malaria genome manually annotated to GO (lots of ISS updated to IDA, especially for cellular component)
- annotations will be released when genome paper is published
- now working on *T. brucei*
- life cycle ontology in progress (Matt & John Richter will try to speed up DAG-Edit — was very slow because of many many relationships)
- for *S. pombe*: Data now in GeneDB (replaces PomBase)

EBI “GOA”

- see handout; highlights:
  - annotation file releases since last meeting:
    - 5 gene_association.goa_human releases
    - 3 gene_association.goa_sptr releases
  - want GO annotations associated with EMBL-Bank records by end of 2002
  - manuscript submitted to Genome Research
  - possibility of SIB-based SP curators using GO to be explored (SP/EMBL retreat coming up late Sept)
  - UniProt Consortium [SP (EBI and SIB) + PIR] grant funded; will allow more manual assignment of GO terms to TrEMBL entries

TIGR

- two handouts: 1 on eukaryotes, 1 on microbes; highlights:
  - sharing *Arabidopsis* annotations with TAIR
  - Manatee tool: interface for editing GO terms and evidence
  - have RefSeq gi number —> GO ID; GO group recommends using protein id instead (gi's not shared by 3 collaborating nucleotide sequence dbs)
  - 7 microbial genomes annotated to GO; *Vibrio cholerae* on GO site; others awaiting genome completion and/or publication
  - GO terms displayed on CMR

** **action item 2**: TIGR to provide protein id —> GO ID

** **action item 3**: TIGR to send IEA annotations to GO for genomes not sequenced at TIGR

Compugen

- GO annotations updated (August 2002)

Incyte

- academic subscriptions to *PD databases
- Lisa seeking to offer financial support for GO meetings

Action Items from CSH meeting (May 2002)
(also see complete list in appendix 4)

1. Many AmiGO items — see software section

2. Check over GO.xrf_abbs file — essentially done, except for incremental updates

3. GO content: modification vs. biosynthesis — done

4. GO content: evaluate sensu terms — done; essentially all will be kept; more “sensu” terms will be added, as will more generic terms as parents for “sensu” terms

5. GO syntax: use of 'and' and 'and/or'; 'or' in gene associations? — Jane is working on removing most terms with “and”; nothing done with gene associations yet

6. Expansion/clarification of GO documentation — not done yet, but Cath presented a plan of action that sounded good

** **action item 4**: Cath will update documentation and circulate drafts

7. Ontology integrity checking — in progress; one thing done so far is that Amelia has a script that checks for several errors; John has set up SourceForge tracker for suggesting checks

8. Submit GO-slim scripts/rules — ongoing

9. GO-slim naming conventions — was done even before it became an action item

10. DAG-Edit/GOET automatic recognition of ID prefix — not discussed; probably not done yet

11. division of 'part-of' into multiple relationship types — not even started

12. GO dictionary — done, with procedure in place for incremental updates (didn't touch on whether it'll be implemented in DAG-Edit or GOET, or, if so, when)

13. Cross-product tool — nothing beyond current DAG-Edit yet (so cross-products can be done but not as easily as we'd like)

14. New documentation for making cross products in DAG-Edit — not done yet

15. comment field: obsoletes & syntax (GO) — done

16. concurrent assignments: QuickGO, database — documentation on QuickGO at http://golgi.ebi.ac.uk/ego/manual.html and http://golgi.ebi.ac.uk/ego/index_internal.html; Evelyn will try to track down more; nothing on GO database side yet

** **action item 5**: Evelyn to continue tracking down info on QuickGO concurrent assignments

** **action item 6**: consortium, especially Chris M, to revisit concurrent annotations in GO database

17. clustering sequences annotated with GO; tool — nothing yet

18. Short descriptions of IEA/ISS methods — in progress

19. gp2protein file documentation — Amelia did a small amount a while back; no word on updating or expanding it

20. monthly release notes — in progress; see Documentation section

21. monthly diffs — in progress; see Documentation section

22. update to current GO home page make links to Gavin's source — not done yet

23. post DAG-Edit user notes — done

24. GO FAQ — Cath and Rama will work on FAQ (not much done yet)

(other action items were related to organizing meetings)

Ontology Structure & Representation: guest presentation by Chris Wroe, with input from Robert Stevens

I'm not going to try to reproduce Chris' talk (!) but here are some highlights:

- ontologies for biology (such as GO) are best done by biologists for biologists

- description logic systems such as DAML+OIL provide a mechanism for building and maintaining ontologies (easier to maintain consistency and completeness with “hand-crafted” ontologies)

- examples from GONG: finding inconsistencies and missing relationships that would be really hard to find manually
  - used MeSH chemical terms
  - missing 'isa' relationships added
  - some 'isa' relationships made more specific
  - errors corrected (e.g. a 'catabolism' term under a 'biosynthesis' parent)

- DAML+OIL can be used at any point along spectrum — don't have to have formal structures already in place to convert

- OilEd tool now available; previously tools for use with DAML+OIL underdeveloped

- definitions (in the DAML+OIL sense): formal definitions for concepts are easy to create for some (e.g. metabolism) terms but much more complicated for others (e.g. enzymes)

- conversion to DAML+OIL will mean a large increase in source code; difficult or impossible to do DAML+OIL diff; Michel Klein developing “virtual cvs”

- case study (how apt!): medical vocabularies and the “exploding bicycle” — highlights need for constraints on what can be combined in cross-products

Linda Hannick took good notes on this talk, so I've included them as Appendix 2A.

GOAL (GO Annotation Language) update: presentation from Bernard

Once again I'm not going to reproduce the whole presentation. Highlights:

- concept of “structure,” in this context referring to any physical entity, such as a gene product or a cellular component

- structure provides activity

- activity changes structure

- word count on GO terms:
  - most frequently used words are connectors ('of', 'and', 'sensu', etc.)
  - 50% occurred only once; of these 65% are “structure” words

- activity = change in structure over time, or

  A = (delta S)/(delta T)

- defining structure (S) and measuring S and time (T) provides information on the activity (A)

  A(r) —-> A(p)
  activity
  S(1) —-> S(2)

  where A(r) and S(1) are starting activity and structure, respectively, and A(p) is activity provided by new structure S(2)

- concept of “housing structure” S(H) — the nearest common parent of S(1) and S(2), not affected by the activity that converts S(1) to S(2); relevant to measuring time (T) — relative, not absolute, time is what's important

  A = [S(1)S(2)]/S(H)

  can compare different activities using function S[S(1)S(2)]

  A = S[S(1)S(2)]/[TS(H)]

- collaborating with Rex Chisholm to try this for Dicty; update later

Linda's notes are included in Appendix 2B.

Database & Software Issues

DAG-Edit & GOET

- John hopes not to do any more development on DAG-Edit. There are a few bug fixes outstanding, but he won't add new features. John would like someone else (a Java programmer) to take over DAG-Edit maintenance; Sue offered to ask Danny Yoo to do it.

- John will add an integrity check to the flat file helper to check for deletion of terms that were present in the files loaded.

** **action item 7**: add check for term deletion to flat file helper

** **action item 8**: Sue will ask Danny to take over DAG-Edit maintenance

** **action item 9**: Amelia will collect bug reports and feature requests from curators. If John can't act on feature suggestions, perhaps Danny can.

- John is developing GOET in the context of image annotation for Drosophila. This takes his time away from GO in the short run, but he will be working on the infrastructure of GOET, which will eventually benefit GO.

AmiGO

Brad has made progress on most of the AmiGO-related action items from last time:

a) make display of NOT data possible/correct in AmiGO (e.g. FBP26 for SGD; FlyBase, others have more) — DONE

b) metareference for curator refs for AmiGO (BDGP and/or GO): create a metareference for linking for curator refs for definitions for AmiGO (e.g. GO:mah, SGD:krc, etc)

Not done yet because there was no way to distinguish a definition dbxref from any other dbxref (also relevant to item l), nor was there any way to tell a reference to a person apart from a reference to a database entry. We'll introduce a prefix to be used for references to curators (GOC:) and Brad will generate web pages to be used as the metareferences.

** **action item 10**: change prefixes to “GOC:” for definition references that represent an individual curator or group of curators

** **action item 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

c) linkouts in AmiGO to sequence in cases such as ISS with ________) — DONE

d) Incorporate GO-Slim scripts into AmiGO — not done yet

e) display comment field in AmiGO — This requires comments to be stored in the GO database, and will be done as soon as they are.

** **action item 12**: Chris to get comments into the database

f) show concurrent assignments in AmiGO — another one in the pipeline, pending addition to GO database

g) add a SourceForge site for AmiGO bugs/requests — DONE

h) gray out obsolete terms (post meeting addition) — DONE

i) link from treeview page to graph view — DONE

j) search function for the comments — again, depends on having comments in database

k) don't automatically toggle to gene product when the search result comes up null — DONE

l) need to make sure that definition references go up with the def, not in the general dbxrefs — can be done once definition references are distinguished from other dbxrefs in the database

m) add ability to upload files for multigene search — DONE

n) GOST, request for it to accept a seqID — programming done; will be "live" once new Linux cluster is installed (probably is by now)

o) want to be able to search with SwissProt accession numbers (this requires a gp2protein file for every organism, nothing for TIGR, PomBase, etc.) — notes aren't quite clear; doesn't seem to be done yet

p) having a way of hiding/deselecting GO terms in BLAST report that you don't believe — hard, and not done, but one can now choose a cutoff score

GO-Slim Issues (overlap between software & annotation):
Many users have asked the model organism DBs (especially SGD) to provide files with gene symbols and GO (or GO-Slim) terms that have been assigned to the gene product. After some discussion we decided to do so, and to include both annotations to the "unknown" terms and genes that have not yet been annotated (the latter will be listed as "unexamined").

Mike Cherry also suggested a table showing each GO term and a list of gene products annotated to it (originally suggested to Mike by Fritz Roth). No decision on this one.

On a related note, Chris has devised, and Matt has tried, a clunky method for generating pie charts using a GO-Slim of one's choosing. The clunky bit is that associations between gene/gene product IDs and GO-Slim terms have to be reloaded into a new database.

** **action item 13**: Add a link to the GO-Slim directory to the home page.

** **action item 14**: DBs to send GO-Slims and lists of all genes to BDGP.

** **action item 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.

** **action item 16**: Add hyperlinks to the gp2protein files: link from web page and from each gene_association file.

Content Issues

How to coordinate work of several curators, geographically dispersed and having backgrounds in different areas of biology, and maintain the consensus-building approach that has worked so well for us?

We agreed that dividing up work based on areas of interest/expertise is a good way to go. To facilitate it, we'll need to keep track of who's working on what. We'll set up "interest groups" for any areas within the ontologies that are likely to require extensive additions or revisions, or to have proposed changes crop up frequently. Curators can join or leave groups as they please. Proposed changes relevant to an interest group should be handled (or at least seen) by that group.

Can we come up with a way to tell whether a given area within an ontology has been extensively reviewed? There was an unfortunate incident recently where a change was made to a bit of the newly revamped 'development' portion of the process ontology. We'd like to avoid this sort of fumble in the future, but it's impossible to tell just by looking at the ontology which bits have been reviewed thoroughly and which parts still look much as they did two or three years ago. There's a lot of information socked away in CVS log files, the email archive, and meeting notes, but it would be much more convenient for curators if the excavation of ontology content history could be streamlined.

In the long run, it should be possible to flag terms as "reviewed" in the database, but there's no simple solution for the flat files. We'll just have to keep records as well as or better than in the past, and spend time and effort to keep each other informed. It's not hopeless, though; there are a couple of things we can do to facilitate communication and record-keeping.

To help with record-keeping, all ontology content changes will be put in the SourceForge curator request tracker from now on. Jane and Midori can add any GO curator to the list of possible assignees; every member database whose curators have GO CVS write access should have at least one curator on the SourceForge list.

Note that putting things in the SourceForge tracker is not mutually exclusive with sending messages to the GO list. Any item that obviously is, or might be, involved or controversial should still go to the list. Err on the side of sending more things to the list if it's not clear.

To keep everyone informed, we'll run a script that extracts the summary lines from new SourceForge entries and emails the resulting list to the GO mailing list. (In theory anyone can join the mailing list specifically for the SourceForge tracker, but few will want to, because the volume of email is huge and most of it is administrative dross.) Anyone can then follow the discussion of any item that looks interesting (try the SourceForge "monitor" option — it's cool!), and anyone can choose to take the discussion onto the GO mailing list.

We decided that there is no need for a "GO curators" mailing list: "interest groups" are likely to change over time, and anything of relevant to more than the interest group should go to the main GO mailing list anyway.

** **action item 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).

** **action item 18**: All content changes, no matter how small, should go into the SourceForge tracker for archiving purposes. Summary entries should be nice and informative.

** **action item 19**: Set up script to email summaries from new (open) SourceForge tracker entries.

On the specter of excessive granularity (a long involved discussion indeed):

We reaffirmed that gene products should not appear as concepts (i.e. as ontology terms). But under some circumstances it is acceptable to mention gene products within ontology terms. The issue to be resolved is how fine-grained we should be in children of “protein biosynthesis,” “protein binding,” and some others.

Many of the children of “protein binding” and of “protein biosynthesis” mention specific individual proteins; see the MGI handout for a list of terms that have come into question.

There is an additional concern with protein biosynthesis terms: many of the too-specific ones added recently are actually intended to capture the results of experiments that measure levels of specific proteins, but do not distinguish effects on translation (the restricted definition of “protein biosynthesis,” which is what we use in GO, and have implicitly decided to keep using) from effects on other steps in the overall process of making a protein (e.g. transcription, modification).

We thought that adding terms for binding to (or biosynthesis of) any specific protein was reasonably consistent with the logic we apply when considering new terms, but we questioned the utility of having many many very specific terms.

We agreed that we would keep or add terms that represent different mechanisms, such as “covalent protein binding” and “non-covalent protein binding” (hypothetical examples) or “viral protein biosynthesis.”

Michael came up with a two-part test; we can keep/add a “protein X biosynthesis” term if both criteria are met:

1. There is something specific about the biosynthesis of protein X, i.e. there are gene products involved in X biosynthesis but not general protein biosynthesis.

2. The proposed term is not redundant with any other process term. For example, we will make “glycoprotein biosynthesis” obsolete because it is redundant with “protein glycosylation.”

The same test can be applied to binding, transport, etc.

But how to avoid losing information? Curators often want to capture what is known, as when an experiment detects binding to a particular protein substrate or altered levels of a specific gene product.

The coffee break “Round Table” discussion led to a proposal: eventually make children of “protein binding” obsolete, and instead use annotation to indicate which protein is bound by the gene product of interest. The annotation would use the generic “protein binding” GO term, and a new column in the gene_association file where we can store an ID for the protein that is bound.

Inevitably, though, there's a catch: the world is not yet ready for us to implement this in all situations. If the gene product being annotated binds a class of proteins — the example was actin — rather than a single protein, we're SOL for the present. In time there will be UniProt IDs representing protein families, but that could take months or even a year or two. There was some discussion of what to do in the meantime; the conclusion was to apply a couple more tests to identify terms that we should keep for now but make obsolete later. First, check over annotations that use the term; second, check whether the term has any children. Annotations will help us figure out whether the term meets the first criterion of the two-part test. A term that has children is most likely a useful grouping term.

The same considerations, and possible future solution, apply to “protein X biosynthesis.” To address the issue of experiments that detect changes in levels of a particular protein, we have decided to consider adding terms for “gene expression” and regulation of same, but further discussion is required before we add them (I suspect that counter-arguments will be raised). If they are added, the new gene_association column could be used with them in the same way as proposed for protein binding.

** **action item 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.

** **action item 21**: Circulate a proposal for incorporating “gene expression” and “regulation of gene 77expression” terms and definitions.

** **action item 22**: Discuss this again at the next meeting!

<u>“Cellular process”</u> to distinguish from multicellular processes was generally well received. Examples where the distinction would be useful are cellular morphogenesis vs. organ or body morphogenesis, cellular respiration vs. breathing, etc. It will take some work to define “cellular process.”

** **action item 23**: Propose definition for “cellular process” and discuss on mailing list.

** **action item 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 pr7ocess.”

“Cell surface” and related terms: these were added recently by TAIR curators, to capture information from experiments in plants that can narrow down localization to plasma membrane or cell wall but can't distinguish between the two (that's what's meant by “cell surface” in plant literature). The definitions and placement of the cell surface terms were discussed, and changes recommended.

We also discussed other cellular component terms in the area of external or surface structures such as cell walls. The fairly generic term “external protective structure” will be changed because “protective” sounds too much like a process; we came up with “encapsulating.” The revised term, “external encapsulating structure,” will become a child of extracellular. The definition should mention that the structure lies outside the plasma membrane and surrounds the entire cell. We should also review the cell wall terms to make sure they're placed correctly — apparently the plant cell wall term should be under extracellular.

One thing that came up is that there are no cellular component terms that really reflect boundaries (as opposed to physical parts) such as that between inside and outside the cell. It will be interesting to look into boundary terms, considering how they might be defined and where they might fit relative to existing terms.

** **action item 25**: TAIR curators to improve definitions of “cell surface” and its children.

** **action item 26**: Change wording of GO:0030312 to “external encapsulating structure.” Circulate new definition; make sure Michelle Gwinn has a chance to comment.

** **action item 27**: Review all “cell wall” terms to check parentage. Plant cell wall does need to be moved.

** **action item 28**: Start thinking about terms (and definitions, of course) to capture concept of boundary.

Transport terms: Dianna Fisk (SGD) is collaborating with Can Tran, who works on TC. Function terms will thereby be kept consistent with what's in TC. Most transport process terms should be OK, but as always any problems should be noted and sent to the list. Transport terms that mention specific proteins should be put to the same test as binding and biosynthesis terms (see above), although we expect that the results will prompt us to keep more of the transport terms.

Susceptibility/resistance: We decided to make all terms that say “X susceptibility/resistance” obsolete because they really represent traits. The biological processes that we were trying to represent can all be covered by “response to X” terms (many of which already exist; others can be added).

IDs: should we encode F/P/C in the GOID? Although some users have asked for this (for convenience), the overwhelming consensus was that we will not add anything to current GOIDs to show whether the term is molecular function, biological process, or cellular component. We will eventually be in a position to build links between what are now the three separate ontologies, so it's better to use a single ID space for them.

Annotation Issues

We receive frequent requests for GO terms/IDs to be associated with UniGene IDs. One way it can be done is via a UniGene <—> LocusLink file available from NCBI.

** **action item 29**: Create UniGene <—> GO file (Daniel)

Issue raised by TIGR (Linda Hannick): how to represent annotations made using multiple BLAST hits or similarity to a domain or family (rather than similarity to one other gene product)

The problem: they feel that they're losing information about the annotation/curation procedure by putting only one accession number in the “with” column. For many of these comparisons, several sequences have to be included, and the similarities among them taken together, to get a believable conclusion about the annotations for the gene product of interest. Furthermore, many of these curated sequence sets are not yet published.

Discussion centered mainly on whether the situation was best covered by using ISS or IC as the evidence code. The eventual decision was to continue to use ISS.

Some key points that came up in the discussion (documented for posterity):

- The argument in favor of IC was that considerable curator judgment is involved in making the determinations, which makes the procedure different from simply running BLAST and looking at the best hit. There was concern about “polluting” ISS by including cases where similarity is to a family rather than to a single gene product.

- The counter-argument was two-fold. One point is that multiple sequence alignments are nevertheless still analyzing and comparing protein (or nucleic acid) sequences, and most curators have been mentally including these analyses under “ISS” all along, viewing them as consistent with the currently defined scope of ISS.

- The second point was that IC is used in a well-defined set of circumstances, for a well-defined purpose. It would “pollute,” or at least confuse, the scope of IC to use it for annotations that are based on sequence similarity; also, one could follow similar logic to broaden IC to include all curator evaluation of experimental results. We decided not to relax the current definition and scope of IC.

Conclusion: allow >1 entry in “with” column for ISS Curators then enter any accession numbers available, and include an ID that allows a link to a page describing the entire set of sequences used.

** **action item 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

** **action item 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)

Pseudogenes and other “doubtful” genes: If a gene is known to encode an RNA or protein product, there's no doubt that the product(s) can be annotated with GO terms (or the gene can be annotated in lieu of direct gene product annotation if necessary). Genes that look as though they encode a product (e.g. open reading frames with no stops) but haven't been individually studied tend to be annotated. If something is unmistakably a pseudogene — lots of frameshifts, etc — it's not annotated.

But what about other cases that fall between the “obviously OK to annotate” and “obviously pseudogene” ends of the spectrum? From Michelle Gwinn:

> We have a class of genes which according to our sequence data have either a single frameshift or a single stop codon in their coding sequence. However, they also have screaming good hits to other characterized proteins and to HMMs that span the problem in the ORF. We reflect the presence of the defect with an addition to the common names of the proteins.

The concern is that a single frameshift or stop may be read through, or could even reflect a sequencing error. To avoid losing information, we've decided that the best way to handle these cases is to use SO annotation to document the frameshift/stop/whatever anomaly, and GO annotations to capture what the product is thought to do if it is indeed expressed.

Shared annotations: For some organisms, gene products are annotated by more than one group (e.g. MGI and SWISS-PROT do mouse; TIGR and TAIR do *Arabidopsis*). We must avoid circular annotations, especially those based on sequence similarity (ISS). Most (all?) of the groups that inherit annotations from another source tag them in the gene_association file some way. For example, MGI has a special reference used for annotations inherited from SWISS-PROT. This was regarded as a good way to handle shared annotations; any group that doesn't do something of the sort already should adopt the practice.

** **action item 32**: Each group that shares annotations should tag the ones that come from the other group(s).

** **action item 33**: Document this decision, and how to implement it.

Documentation Issues

Monthly logs: Amelia has been working on a script to detect differences between one version of GO (ontologies + definitions); she showed sample output that was very well received. There is still a bit of work to do to get it to prime-time quality, but it is in very good shape. We will run the script every month, when the flat files are archived and database releases made. In addition to running it regularly, we'll include it in the software repository on SourceForge, so that anyone can run it to compare any two versions of GO.

** **action item 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.

FAQ: Chris will help Cath and Rama set up a FAQ-o-matic page; thereafter, anyone can enter question and answers. Cath and Rama will do a bunch to get things started and make sure the FAQ covers questions that we already know crop up frequently.

** **action item 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.

** **action item 36**: EBI GO curators circulate a set of instructions for using CVS.

Other Items of Interest

GO <—> UMLS: Jane will visit NLM for about a month starting Sept. 15. She will learn all about UMLS, and help them incorporate GO into the "Metathesaurus." That is, GO will become one of the ontologies indexed in the metathesaurus. Jane and some NLM people have already done a test integration. MeSH terms will be reviewed and new ones added in light of indexing GO in UMLS. Jane will report on this work at the next meeting.

Funding: For the NIH grant, there's a progress report due soon (December 1?). Judy will coordinate, and email anyone who should contribute material.

We will apply for five years when we renew; the renewal is due March 1, 2003. Judy will also coordinate this. There will be four aims:

1. Develop and support ontologies for molecular biology.

2. Annotation using ontologies for informatics systems of consortium members; this will include support for meetings.

3. Provide informatics resource; covers database instantiations, data repository and means of access, and software tools.

4. Outreach: support for ways to provide training for new groups starting to use GO, perhaps by having them visit a "GO site". A "visiting scientist" sort of thing could also be a good way for GO curators to take advantage of domain experts' knowledge. Meeting support might also fall under this aim.

Aims 1, 2, and 3 are essentially the same as in the original grant, with the scope of Aim 1 expanded a bit. Aim 4 is modified from the original aim to have other database groups join the consortium.

We would also like to support an effort to annotate bacterial genomes i.e. those not already done or in the works at TIGR) using GO. *E. coli* and *B. subtilis* are the most obvious ones; genomes sequenced at Sanger would also be good.

** **action item 37**: Progress report for current grant.
** **action item 38**: Prepare renewal grant application.

GOBO: Covered in Michael's talk at the Users meeting. We have a supplement to the NIH grant to fund work on SO; Suzi will hire two people, one more biology-oriented, the other more techy, for a year.

SOFG: conference coming up in November.

Web pages: We'll keep the current appearance for the time being, but that shouldn't stop us form improving the organization. The home page can be split into a few shorter pages, based on the work Amelia did earlier.

** **action item 39**: Prepare a site with mock-ups of GO web pages derived by splitting up the current home page sensibly.

Next Meetings:

The next Consortium meeting will be January 25–26, 2003 in St. Croix. Plan to arrive on Jan. 24 and leave on Jan. 27. John will make a group reservation; when we get the email about it, we must act promptly because rooms will go fast. There won't be a Users meeting.

After that, the next meeting will be hosted by TIGR in June 2003, with a Users meeting. Linda will check on available dates; our first choice is June 2–4 (users on Monday June 2, consortium Tues–Wed June 3–4). Alternate dates are June 18–20.

Appendix 1: Collected Action Items (numbered in the order the appear in the main document)

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

2. TIGR to provide protein id —> GO ID.

3. TIGR to send IEA annotations to GO for genomes not sequenced at TIGR.

4. Cath will update documentation and circulate drafts.

5. Evelyn to continue tracking down info on QuickGO concurrent assignments.

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

7. Add check for term deletion to flat file helper.

8. Sue will ask Danny to take over DAG-Edit maintenance.

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.

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

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.

12. Chris to get comments into the database.

13. Add a link to the GO-Slim directory to the home page.

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

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.

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).

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

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

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.

21. Circulate a proposal for incorporating “gene expression” and “regulation of gene expression” terms and definitions.

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

23. Propose definition for “cellular process” and discuss on mailing list.

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.”

25. TAIR curators to improve definitions of “cell surface” and its children.

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

27. Review all “cell wall” terms to check parentage. Plant cell wall does need to be moved.

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

29. Create UniGene <—> GO file (Daniel)

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.

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).

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

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

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.

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.

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

37. Progress report for current grant.

38. Prepare renewal grant application.

39. Prepare a site with mock-ups of GO web pages derived by splitting up the current home page sensibly.

Appendix 2: Linda Hannick's notes on presentations by Chris Wroe and Bernard de Bono at the GO Consortium meeting, 10 Sept. 2002

A. DAML+OIL
Chris Wroe / Robert Stevens

Experiments in how you can use hand-crafted text → software-based technology
What we can do, not tutorial…
Helen Parkinson

- **What does the technology offer?**

Process:

- Electronically generate rather than add manually.
- Pathway; not all or nothing; some benefit part way too…
- Simple additions from yesterday

  Making relationships to additional parents, etc

  Finds biological content error(s), finding relationships that are problematic

  Suggests additions, finds the missing relationships that are very hard to find by hand, suggests additional.

  Inconsistencies reasoned out (e.g., a case of catabolism under biosynthesis.)

- What software is available?

GONG: what have we done so far?

Developing a stepwise methodology
Incremental migration path adding semantic content to the GO *in situ*

1. Syntax transformation to DAML+OIL
2. Reasoning over existing content
3. Adding partial concept descriptions
4. Adding complete " "
5. Concept composition at the point of use

Allow the creation of new ontology terms at the point of use.
→Isa has to be done manually; P is easy
less hard work than doing it all by hand.

Migration path

Definitions/descriptions (carbohydrate metabolism) broken down to a DAML+OIL *necessary and sufficient conditions.*
Complete definition: biosynthesis of an amino acid
Natural language pulls out the essentials
Natural lang tool what *you see is what you meant*
Metabolism terms easy; enzyme terms very complex to describe
Absolutely explicit; lots of restrictions onProperty and has-class restrictions
Top-down approach would be easier (dehydrogenase defined before malate dehydrogennase)

Scripts were central

Used as much automation as possible
Many term phrases fit a stereotyped pattern
Metabolism for example
Hard coded
UMLS lexical normalization tools to match up concepts from different ontologies

May also help the parsing task

- Additional DAML+OIL definitions represent a significant increase in the amount of 'source code'
- Introduces large numbers of interdependencies (lots of these were missed in the hand-built ontology.
- Michael Klein – conceptual cvs for DL
  Can't just do a diff on DL; need a cvs. Meeting tomorrow. Will e-mail the group re this.

Software

- DL datastructure with API
- Editor gui OilEd
- Ontology server

Case study from forerunners in medicine (SNOMED)

- Learn from their mistakes; already avoided mistakes of early medical terminologies
- Similar to medicine; large, complex concepts
  - o SNOMEDrt relational terminology
  - o 200K-300K concepts at the present time
  - o results 200K concepts dissected over 2-3 yrs by 9 half-time clinicians (double coverage)
  - o ~20M investment
  - o tried to use scripts and tools; propagation of concepts like we are discussing
  - o major early benefit was a more complete taxonomy for accurate retrieval of records
  - o not open source; there is a gathering force behind going open source (Richter, Chris Schut) (global technological project, GTP).
  - o Formal def of terms useful resource in its own right irrespective of DL reasoning

- But different; well specified use—annotation
  Relatively small group of people who are highly skilled
  Medical record keeping more for the accountants

What additional software is necessary?

BioOntologies people using DL? Not extensively
2 diff ways to use
as a standard, because it works with ontologies
with property-based descriptions

Open source? OIL is; Java client pops it in (Robert) License for display is _36000 on Solaris.

Problems: Scaling
The combinatorial explosion
Example Burns
How expensive Read II grew from 20K to 250K terms in ~100 staff-years, but still too small to be usefule
But too big to use…

(SNOMED 3.5) Beat the explosion by having ~12 separate taxonomies, the elements of which can be combined to form more complex concepts.
Didn't work because the sensible options have to be defined in the user interface.

- o No grammar rules
- o Possible to make nonsense terms
- o Impossible to detect equivalent terms, or classify composition

Need a reference terminology in the middle.

B.
GOAL (Bernard)

It is structure that provides activity.
It is activity that changes structure.

Organism bias is a cumulative effect of structure bias that has infiltrated GO.

Word counts in GO
Split into two sets, A and B.
Of, and, etc in A
Set B

> (90% of words used in GO terms).
> Occur 10 times or less in the DAG.

>50% of B set occur only once throughout DAGs

> 65% of set B are physical objects in our universe (glutamate)

three DAGS alike
<u>most of the words are structures.</u>

$$A=\frac{\text{change in structure}}{\text{change in transitionTime}} \qquad \frac{\Delta S}{\Delta T}$$

Is it possible to have any sort of value for ΔS and ΔT?

Having a handle on Structure will have a profound effect on Activity

Hypothetical structure classification:
Small mol
Gene prod
Complexes
Cells
Anatomy

Map activity on graph of above.
2 structures more similar will be closer on graph.A
Can impose different organisms on the graph.

GOAL Object Definition
A_r shifts S_1 to S_2.
A_p is new activity.
Navigate the structural graph by activities.
Distance along the tree will be significant.
$Ar=s(S_1,S_2)/t(S_1,S_2)$
r is “required”
p is “provided”

Extend to any level of complexity.
S_H housing structure :
The first part of the node that S1 and S2 have in common.

$$A = \frac{S_1, S_2}{T(S_H)}$$

Profile comparisons of ACT objects

Now can compare Activities much in the way the BLOSUM matrix is used. Compare whole-genome physiologies.

Show on the same structural graph what you mean by an activity.

Appendix 3: Progress Reports

A. FlyBase Progress Report, Sept 2002.

Cambridge Meeting.

1. GO terms added by continued literature curation of primary papers and personal communications by Cambridge FlyBase curators Rachel, Gillian and Chihiro.

2. Kerry Knight is currently assigning GO terms as part of her clean up of free text in FB; referencing to primary papers.

3. All outstanding SWISS-PROT records (~1000) that were attached to a FlyBase genes have now been analyzed and GO terms added based on the summary comments. GO terms are referenced directly to the SWISS-PROT record. In addition, Eleanor Whitfield at SWISS-PROT is assigning GO terms to new SWISS-PROT records, and SWISS-PROT records updated from SpTrEMBL. These are referenced to papers listed in the SWISS-PROT record and are also incorporated into our files. We now periodically do a check to ensure that all relevant SWISS-PROT entries are curated.

4. Becky is currently curating recent reviews, mainly on processes e.g. oogenesis, embryogenesis, organogenesis, signaling etc. to increase the number of process GO annotations in FlyBase.

5. Work is ongoing to increase the number of definitions for fly-specific GO terms especially for embryogenesis terms.

6. We have received a file of predicted GO annotations from the FB/PANTHER collaboration. A paper describing this experiment has just been submitted to Genome Research. The predictions have not been parsed into FB. The reason is that this analysis will be redone on the new Release 3 sequence.

7. The next major task will be to re-annotate for GO terms the Release 3 protein set. That should keep us busy for some time.

Rebecca & Michael.

B. Gene Ontology Annotation @ EBI

The GOA Project is headed by **Rolf Apweiler**
GOA Annotation Coordinator: Evelyn Camon (camon@ebi.ac.uk)
GOA Electronic Coordinator: Daniel Barrell (dbarrell@ebi.ac.uk)
URL:http://www.ebi.ac.uk/GOA

Last Updated: 03-SEP-2002

Current Status:
We have made 5 releases GOA Human and 3 release of GOA SPTR(GOA-All) on the EBI and GO ftp sites. In SRS these releases are merged in the one database called GOA. The recent release of all our GO annotation makes SWISS-PROT group at EBI a considerable contributor to the GO consortium annotation effort providing ***over 2.1 million*** GO associations across ***507964*** SWISS-PROT and TrEMBL entries covering ***45407 species***. GOA Human releases are in keeping with our Human Proteomics Initiative and GO Consortium agreement to fast-track functional annotation of the human proteome. We have not yet integrated GO data from other Consortium groups due to lack of manual annotation with PUBMED references in the association files. We are working particularly closely with Mouse Genome Informatics (MGI) and FlyBase group to resolve these matters. As IPI is now indexing Mouse data we will next work on releasing GOA Mouse.

Discussions have been initiated with EMBL-Bank on how to transfer GO annotations from GOA into EMBL flat files via its db_xref. It is decided to add a link from EMBL-Bank flat files directly to QuickGO eg. db_xref="GOA:P22301". It is hoped that this will be achieved by the next EMBL release, which will be made public in few weeks time.

EBI maintains SWISS-PROT keyword 2 go and InterPro 2 go mappings these are updated on a regular basis and shared with the GO Consortium where they have been used to enhance their data sets as well as those of external GO users (Microarray/mass spec). We are also working closely with PIR to help their keyword mappings.

A GOA paper has been submitted to Genome Research.

The GOA project is ahead of schedule on all its grant deliverables.

HOW IS GO ANNOTATED IN SWISS-PROT/TrEMBL/InterPro/?
GOA is produced by electronic and manual efforts

The large-scale assignment of GO terms to SWISS-PROT and TrEMBL entries involves electronic techniques. This strategy exploits existing properties within the entries including the presence of keywords and Enzyme Commission (EC) numbers as well as the presence of cross-reference to InterPro entries, which are manually mapped to GO.

Electronically combining these mappings with a table of matching SWISS-PROT and TrEMBL entries generates a table of associations. SWISS-PROT keyword and InterPro to GO mappings are maintained in-house and shared on the GO home page for local database updates.

Manual assignment of GO terms by SWISS-PROT curators uses published literature and provides more reliable GO annotation. On each release of GOA, annotation with electronic evidence codes (IEA: 'inferred from electronic annotation') will be replaced with associations using codes that imply more experimental evidence.

RETRIEVING DATA FROM GOA

There are various ways of accessing and searching GOA project data, including several web-based browsers. The GOA files can also be downloaded.

Resource	Description
Web-based tools	
QuickGO	A fast web-based browser with access to core GO data and up-to-date electronic and manual EBI GO annotations. URL: http://www.ebi.ac.uk/ego/index.html
SRS	Search the GOA database or a mirror of the GO consortium repository (GO). URL: http://srs.ebi.ac.uk/
Proteome Analysis Pages	GO annotations have been produced for classification of proteins belonging to each complete proteome. On the Proteome Analysis Pages a slimmed down version of GO (GO-slim), representing high-level GO terms, is displayed as a proteome overview. URL example: http://www.ebi.ac.uk/proteome/HUMAN/go/go.htmlEBI's GO-slim see: http://www.ebi.ac.uk/proteome/goslim_terms.html
InterPro	GO annotations made by InterPro are visible directly in InterPro entries. URL example: http://www.ebi.ac.uk/interpro/Ientry?ac=IPR000402
AmiGO	GO Consortium browser with access to core GO data and released GOA data. URL: http://www.godatabase.org/docs/docs.html
Downloads	
GOA 'Association File'	This is a tab-delimited file of associations between gene products and GO terms and is the most common form of data transfer within the GO Consortium. For more information on our format read the GOA README file (http://www.ebi.ac.uk/proteome/goa/goaHelp.html) Two separate GOA association files are currently produced. Human GOA file access (contains GO annotations for all proteins in the nonredundant human proteome set): ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz http://www.geneontology.org/gene-associations/gene_association.goa_human SPTR GOA file access (contains GO annotations for all proteins in SWISS-PROT and TrEMBL): ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/SPTR/gene_association.goa_sptr.gz http://www.geneontology.org/gene-associations/gene_association.goa_sptr
GOA Xref File	For each GOA release we also distribute a file of cross references that displays the relationship between the entries in the GOA data set with other databases, such as EMBL/Genbank/DDBJ nucleotide sequence databases, HUGO and LocusLink and Refseq. GOA xref file: ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/

STATISTICS

Statistics for GOA-Human and GOA-SPTR association files are available from the GOA homepage. (http://www.ebi.ac.uk/GOA)

GRANT SUPPORT:
GOA is supported by Grants QRLT-2001-00015 and QLRI-2000-00981 of the European Commission and a supplementary NIH grant, 1R01HGO2273-01.

CONTACTING GOA:
Post:
EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus
Cambridge
CB10 1SD
UK

Phone: +44 (0) 1223 494444
Fax: +44 (0) 1223 494468
E-mail:goa@ebi.ac.uk

CREDITS:

Daniel Barrell – GOA File updates
David Binns – QuickGO
Wolfgang Fleischmann –Automation Coordinator
John Maslen - Talisman
Paul Kersey – Xref file & data set generation
Michele Magrane & all curators - GO Annotation
Nicola Mulder, Alex Kanapin & Annotators – InterPro
Rodrigo Lopez, Nicola Harte - SRS
Midori Harris, Jane Lomax, Amelia Ireland, Cath Brooksbank – GO Curators
Rolf Apweiler –SWISS-PROT Coordinator
Peter Stoehr - Head of Database Operations (EMBL-Bank Issues)

C. MGI Gene Ontology Progress Report Sept. 2002

General:

We continue to focus on extending our goal to have annotation for all genes in the database. Our efforts have focused on three areas:

1. Adding annotation to genes currently without any annotation
2. Replacing annotations that were "fished" from text records with literature based annotation
3. Annotating genes having no go but having rat orthology

We have constructed a dataset that might be used as a "gold standard" to judge the efficiency of various annotation algorithms. This dataset is comprised of genes that have been hand annotated with evidence codes derived from experimental evidence (IDA, IPI, IMP, IGI). A second dataset derived from this series has only those genes that have had the same GO ID applied more than once by any combination of these.

MGI GO STATS as of August 27, 2002.

Annotation Type	**30-Apr-02**	**27-Aug-02**	**Change**	**% Change**
Total Genes annotated:[1]	7600	8576	976	13
Total Hand Annotation				
# of Genes	2125	2646	521	25
Orthology:	19	24	5	26
"IEA"				
SwissProt to GO	4852	6123	1271	26
Interpro to GO	3376	3529	153	5
EC to GO	662	658	-4	-0.6
MLC Scan	40	40	0	0
GO Fish	2337	2228	-109[2]	-5

Beyond GO

The phenotype ontology is continues to be developed with the aid of the DAG-Editor[3], which has facilitated term merging and increasing the complexity of the DAG structure.**Too many unnecessary GO terms:**

On the issues of excess granularity

The GO was originally set up as a vocabulary to describe the molecular function, process, and cellular location of a gene product that could be used across model organism databases. However, recently, the GO appears to be

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

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

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

growing in areas that appear to reflect a cross over between product name and function and process. There are three example areas:

1. Protein Binding
2. Protein Biosynthesis
3. Immune Response : interleukin X biosynthesis…..

1. The function term “:Protein Binding” coupled with the “with” statement is intended to describe the interaction of a gene product with another protein. The creation of dozens of children that specifically refer to a single gene product in a single type of organism (mammal), as in the cases of interleukin-X binding, where X is a specific molecule, unnecessarily increase the granularity of the GO in a species specific manner.

2 and 3 . Protein Biosynthesis was originally meant to describe the processes involved in the formation of a peptide bond, either on the ribosome or not. The creation of specific terms for single instances of proteins is unnecessary. If the term is NOT meant to decribe processes involved in peptide bond formation, it should not be a child of this term. The use of the term “XYZ protein biosynthesis” to be used for a description of any unknown process or combination of processes involved in altering the level of a particular gene product is ambiguous. If there is not evidence to pinpoint transcription, RNA processing, translation, post-translational processing, or RNA and/or protein degradation as the process or processes that are involved in the gene product to be annotated, then perhaps no annotation should be applied. If we proceed down this path, then XYZ biosynthesis will need to have specific children, XYZ biosynthesis, transcription, etc.

1. The first issue begins in protein biosynthesis, where we currently have:

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

***TRAIL receptor biosynthesis +**

translational elongation +
translational initiation +
translational termination +
viral protein biosynthesis

*What we do not need is a separate term for each protein. As I understood from discussions on the GO-list, these terms were intended to encompass everything that goes into making the protein, from transcription, translation, and perhaps even degradation. They are intended to capture experiments that use protein /gene product A to influence the (levels) of protein/gene product B. There may be 100 steps between the two. This is making the GO terms experiment driven rather than the other way around.
Such experiments are just NOT useful as evidence for any GO terms. They suggest experiments to be done.

**The second issue regarding protein biosynthesis is that adding lipids and carbohydrates to proteins is a post-translational modification and does not belong under protein biosynthesis. The term “protein biosynthesis” should be restricted to processes that form a peptide bond, either on the ribosome (mostly) or not (antibiotics).

2. A second area of is the growth of a separate term for each protein binding:.

protein binding [GO:0005515]
alpha-catenin binding
ARF binding
beta-amyloid binding
beta-catenin binding
cadherin binding
calmodulin binding +
clathrin binding
collagen binding
cyclin binding
cytokine binding +

- chemokine binding +
- granulocyte macrophage colony-stimulating factor complex binding
- interferon binding +
- interleukin binding +
    - interleukin receptor +
    - interleukin-1 binding +
    - interleukin-10 binding +
    - interleukin-11 binding +
    - interleukin-12 binding +
    - interleukin-13 binding +
    - interleukin-14 binding +
    - interleukin-15 binding +
    - interleukin-16 binding +
    - interleukin-17 binding +
    - interleukin-18 binding +
    - interleukin-19 binding +
    - interleukin-2 binding +
    - interleukin-20 binding +
    - interleukin-21 binding +
    - interleukin-22 binding +
    - interleukin-23 binding +
    - interleukin-24 binding +
    - interleukin-25 binding +
    - interleukin-26 binding +
    - interleukin-27 binding +

interleukin-3 binding +
interleukin-4 binding +
interleukin-5 binding +
interleukin-6 binding +
interleukin-7 binding +
interleukin-8 binding +
interleukin-9 binding +
cytoskeletal protein binding +
DNA topoisomerase I binding
dynein binding +
enzyme binding +
eukaryotic initiation factor 4E binding
gamma-catenin binding
growth factor binding +
hemoglobin binding
histone binding
HSP70 protein binding +
immunoglobulin binding +
importin-alpha export receptor
intermediate filament binding
ISG15 carrier
KU70 binding
lamin binding
lipoprotein binding +
metarhodopsin binding
neurexin binding
nuclear localization sequence binding
peroxisome targeting sequence binding +
poly-glutamine tract binding
polypeptide hormone binding +
profilin binding
protein amino acid binding +
protein C-terminus binding
protein carrier
protein domain specific binding +
protein signal sequence binding
RAN protein binding
Rho binding +
RPTP-like protein binding
SNARE binding +
snoRNP binding
syndecan binding
TATA-binding protein binding
TRAIL binding
transcription factor binding +
Wnt-protein binding

This loses the utility of the Protein Binding and With fields. Are we going to have a separate term for every single pair of proteins. The chemokine and interferons conceivably could be expanded in a like manner

This is not needed. The primary term plus the "with" field is sufficient. Algorithms could be written where if the pairs are annotated properly, one could search the "with" field to come back with all binding partners.

3. A third area is sort of related to the "biosynthesis " issue again> Why are separate terms for the biosynthesis of each interleukin needed??

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 +

All of these could be easily described using GO terms for translation, protein processing, etc. Again, we do not need a term for each specific protein product. These too appear driven by the desire to want to use an experiment to create a GO term.

We need to decide how granular the GO needs to be.

D. GO Report from SGD

Outline

- SGD Goals for GO Annotations
  - Definitions for GO Terms within SGD
  - Annotations
- GO Tutorial
- GO Tools
- Pathway Tools

SGD Goals for GO Annotations

Definitions for GO terms within SGD
SGD is making a big push to write definitions for all the terms that have been used to annotated SGD genes. There are about 68 component terms, 268 function terms and 287 process terms that need definitions. Each curator writes 2 definitions per month and also if the curator needs to annotate to a term that doesn't have a definition, he/she will write the definition before making the annotation. We are making good progress towards this goal.

Annotations
Our goals for the near future are:
Have at least one annotation for all the named genes. Out of 4297 named ORF's we do not have
any annotation for only 367 loci.
Fill in annotations for genes that have partial annotations.
Polish all the annotations (work on the IEAs and the 'unknown' annotations).

GO Tutorial

SGD has created a GO Tutorial to familiarize users with the Gene Ontology (GO) and how it is used at SGD. The tutorial gives an overview of GO and highlights pages and tools at SGD that use GO annotations with some cool mouseovers. In addition, the tutorial provides links to other sites that may help users take advantage of the power of GO.

GO Tutorial: http://genome-www.stanford.edu/Saccharomyces/help/gotutorial.html

GO Tools

SGD has developed 2 tools to mine GO data. They are the GO Term Mapper and the GO Term Finder tools.

The GO Term Mapper or the GO slim tool maps the granular GO terms used to annotate a list of genes to their more general parent terms (ie. GO Slim terms) from all three ontologies.

The GO Term Finder finds all the terms and their parents for a list of genes (users query). The GO Term Finder gives a tree view of all the terms with the DAG relationships, that the query set of genes have been annotated to.

Both these tools can take a file of gene names or ORF's as input and can be very useful for analysis of expression data.

GO Term Mapper: http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermMapper
GO Term Finder: http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermFinder

Pathway Tools
SGD is in the process of incorporating biochemical pathways into the database using Peter Karp's (Stanford Research Institute, CA) Pathway Tools. A summer student mapped E.C. numbers to metabolic enzymes in SGD (approximately 1000) by using ec2go and searching the literature. In the first build using the Pathway Tools, 828 reactions were created in 163 pathways. We are in the process of refining the pathways. We will be using the E.C. numbers to increase the GO function annotations and hopefully add to the current ec2go file as new GO function terms are created.

E. TIGR eukaryotic GO update September 2002 Linda Hannick

Associations currently at GO:

	Arabidopsis Aug-27-2002
# genes with GO assignments	5089
since last release	798
# terms assigned	10833
molecular function	6564
biological process	2807
cellular component	1462

Associations not yet released to GO

Other euk GO annotations in progress and not yet released include chromosome 2 of T. brucei (manually curated) and O. sativa (IEA).

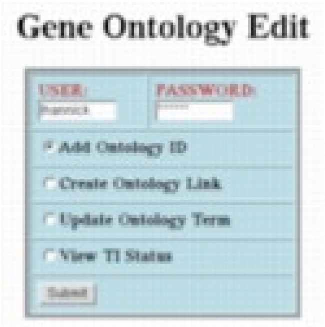
New developments

The Arabidopsis project is now sharing GO annotations with TAIR weekly. TAIR GO assignments will be stored in our database along with our own to prevent duplication of work. They will be displayed on our annotation interface.
We have a new gi2ath association file from GenBank which will be uploaded to the GO ftp site after this meeting.

Software improvements are making it faster and easier to assign GO terms. The Manatee interface now allows editing of GO terms and evidence. A new GO search page allows an annotated search of a particular genome in our database, or a search of a the entire DAG.

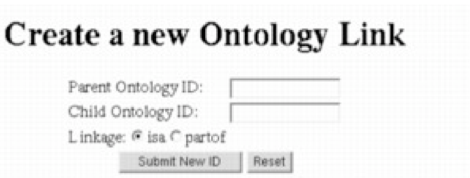
TIGR annotators track new terms using temporary "TI:" ID's. The assignment of temporary terms is now enabled by a set of pages:

The TI: ID's are intended as a tracking device for new terms as they are submitted to Sourceforge. They are replaced automatically in our database as we enter the newly assigned GO: ID, with the TI: ID becoming a synonym to the GO ID in our database.

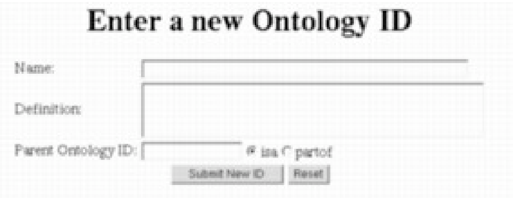

Gene Ontology Edit

USER: hannick
PASSWORD: ******

● Add Ontology ID
○ Create Ontology Link
○ Update Ontology Term
○ View TI Status

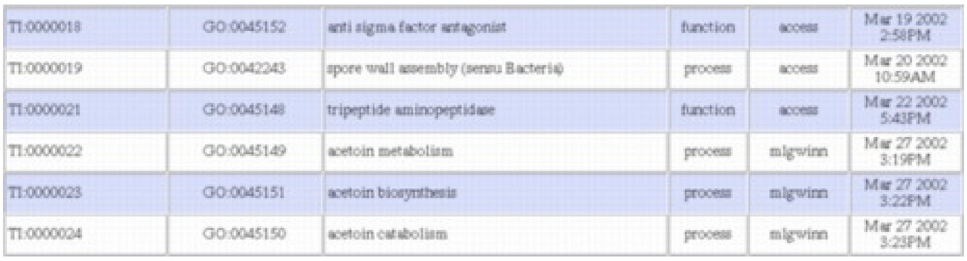

Create a new Ontology Link

Parent Ontology ID:
Child Ontology ID:
Linkage: ● isa ○ partof


Enter a new Ontology ID

Name:
Definition:
Parent Ontology ID: ● isa ○ partof

Track TI ID's



TI:0000018	GO:0045152	anti sigma factor antagonist	function	access	Mar 19 2002 2:58PM
TI:0000019	GO:0042243	spore wall assembly (sensu Bacteria)	process	access	Mar 20 2002 10:59AM
TI:0000021	GO:0045148	tripeptide aminopeptidase	function	access	Mar 22 2002 5:43PM
TI:0000022	GO:0045149	acetoin metabolism	process	mlgwinn	Mar 27 2002 3:19PM
TI:0000023	GO:0045151	acetoin biosynthesis	process	mlgwinn	Mar 27 2002 3:22PM
TI:0000024	GO:0045150	acetoin catabolism	process	mlgwinn	Mar 27 2002 3:23PM

F. TIGR microbial GO update August 2002 compiled by Michelle Gwinn

Associations currently at GO:

	genes	terms
Vibrio cholerae	2924	6243

I just sent a new Vibrio file with more associations, you may have noticed the number went down instead of up, this is due to the removal of GO terms from the plain “hypothetical proteins”, after result of discussion on GO email list.

I also sent a gp2protein file for Vibrio.

Associations (manual) not yet at GO:

Genome	genes	terms
Shewanella oneidensis	3769	8307
Bacillus anthracis	4555	9673
Coxiella burnetii	1467	2711
Methylococcus capsulatus *	2616	4554
Geobacter sulfurreducens *	1916	4078
Listeria monocytogenes*	>1465	>3342 (in progress now)
	------	-------
TOTAL	15788	32665
GRAND TOTAL (with Vibrio)	18712	38908

Genomes pending publication (submitted manuscripts) and subsequent release to GO web page:
Shewanella oneidensis
Bacillus anthracis
(Total of 17980 GO terms)

* indicates annotation is incomplete for that genome, more genes remain from that organism that need to be assigned GO terms

----

Other news:

Our automatic annotation tool is now assigning GO terms to microbial genomes -TIGR genomes, preliminary assignment - followed by manual review prior to release -non-TIGR genomes (IEA) for display on our CMR website
(should we send these to GO?)

Comprehensive Microbial Resource (CMR) displaying GO terms for genes that have them. (should be functional by the time of the meeting)

Rough draft of prokaryotic GO Slim exists, work continues db/software support of GO Slims is under construction

----

If anyone has any questions or wants to chat about any of this, please don't hesitate to email me - mlgwinn@tigr.org

Hope you have a good meeting, see you in the winter,

Michelle

Appendix 4: Action Items from May 2002 Meeting at CSH

ACTION ITEMS FROM MAY 12-13 GO MEETING

Action Item 1 - AmiGO (Brad Marshall, BDGP)
a) make display of NOT data possible/correct in AmiGO (e.g. FBP26 for SGD; FlyBase, others have more)
b) metareference for curator refs for AmiGO (BDGP and/or GO): create a metareference for linking for curator refs for definitions for AmiGO (e.g. GO:mah, SGD:krc, etc)
c) linkouts in AmiGO to sequence in cases such as ISS with ________)
d) Incorporate GO-Slim scripts into AmiGO
e) display comment field in AmiGO
f) show concurrent assignments in AmiGO
g) add a SourceForge site for AmiGO bugs/requests
h) gray out obsolete terms (post meeting addition)
i) link from treeview page to graph view
j) search function for the comments
k) don't automatically toggle to gene product when the search result comes up null
l) need to make sure that definition references go up with the def, not in the general dbxrefs
m) add ability to upload files for multigene search
n) GOST, request for it to accept a seqID
o) want to be able to search with SwissProt accession numbers (this requires a gp2protein file for every organism, nothing for TIGR, PomBase, etc.)
p) having a way of hiding/deselecting GO terms in BLAST report that you don't believe

Action Item 2 - GO.xrf_abbs file (each group): examine the GO.xrfs_abbs file with respect to those abbreviations used by your group, add or submit (to your favorite contact with CVS write permission)

Action Item 3 - GO content: modification vs. biosynthesis (GO) - examine ontologies for consistency of term names in the area of modifications to nucleotides/amino acid residues within the context of an already synthesized nucleic acid/protein

**DONE, except for a few individual cases that aren't straightforward

Action Item 4 - GO content: sensu terms (GO) - evaluate sensu terms, and expand documentation

**in progress

Action Item 5a - GO syntax: use of 'and' and 'and/or' (GO) - evaluate use of 'and' and 'and/or' in GO terms, target for elimination when possible

**in progress

Action Item 5b - possibility of ambiguous gene associations conjoined with 'OR' (BDGP: Chris, John) - discuss possible software solutions to ? of joining two different associations (gene product to GO term) with an 'OR', [NB: resolution of this item was unclear; first communicate with GO people on Action Item 5a and discuss whether there is any real desire/need to do this.]

Action Item 6 - expansion/clarification of GO documentation (GO: Cath B) - Cath will evaluate GO documentation and expand/modify to clarify

Action Item 7a - ontology integrity checking (John) - will create a SourceForge submission page for ontology errors

**DONE!!! 5/13/02

Action Item 7b - ontology integrity checking (each group) - curators should look for ontology errors, and submit them to the SourceForge page that John will create

**two whole entries so far

Action Item 8 - submit GO-slim scripts/rules (each group, as relevant)
- Submit scripts (Chris is fine with Python, or Perl) for using/calculating GO-slims to BDGP

Action Item 9 - GO-slim naming conventions (GO): - confirm/review naming conventions for GO-slims and expand documentation if needed (Michael Ashburner claims that there is a naming convention in the document that he has just written)

**was done already (see go/GO_slims/README)

Action Item 10 - DAG-Edit/GOET (John Richter) - automatic recognition of ID prefix so that one doesn't have to manually change it all the time

Action Item 11 - division of 'part-of' into multiple relationship types (Chris and Jane) - will look into new relationships deriving from the current multiplicity of the meaning of the 'part of' relationship

Action Item 12a - GO dictionary (GO, John Garavelli)- we need a dictionary for John to use for spell checking (John Garavelli wants to write a script for this anyway so he will generate the dictionary)

**DONE; dictionary is updated frequently

Action Item12b - GO dictionary in editor (John Richter) - can write a spell checker for the editor once he has a dictionary

Action Item 13 - Cross-product tool (interested parties (David, Bernard, ?), Chris, and John Richter) - cross-product tool: further discussion will clarify what is actually wanted as well as feasible, so that John can write a plug-in for curators to use via the editor

Action Item 14 - New documentation for making cross products in DAG-Edit as currently exists (GO: Jane, Amelia) - create document on generating cross-products in DAG-Edit

Action Item 15 - comment field: obsoletes & syntax (GO) - move obsolete IDs from synonyms to comment field and institute a regular (as in parsable) syntax for this field

**parsable syntax part is done - syntax established; only thing now is to make sure we use it

Action Item 16a - concurrent assignment protocol/docs for QuickGO (Evelyn) - get documentation from Tom Oinn on how he did it for QuickGO; add to documentation, to explain how this is calculated

Action Item 16b - concurrent assignments from database (Chris) - pull this calculation on concurrent assignments from manual annotations using Database [NB: Fritz Roth is doing some calculations along this line]

Action Item 17a - sequence clustering for sequences annotated with GO (Daniel? Liat?) - take sequences as they are now, run a clustering algorithm, generate trees, attach GO annotations and inspect by hand

Action Item 17b - very cool annotation tool (????, highly dependent on above) - use this to develop an annotation tool that utilizes homology clustering

Action Item 18 - IEA/ISS methods (each group, GO: Midori): Groups to submit to Midori short blurbs on procedures for large scale annotation methods (bulk assignments, particularly with IEA or ISS) with urls to
add to the annotations guide
**I've received ONE response (thanks to Harold Drabkin)

Action Item 19 - gp2protein file documentation (Chris??)- expand documentation for gp2protein files

Action Item 20 - monthly release notes (GO) - take a look at doing monthly release notes
**in progress; item for Sept. agenda

Action Item 21 - monthly diffs (Courtland Yockey) - will investigate DAG-Edit diffs, and communicate with John regarding proceeding further on utility of a plug-in for DAG-Edit that could do this
**progress on item 20 is relevant

Action Item 22 - update to current GO home page (Karen) - make links to Gavin's source

Action Item 23 - DAG-Edit user notes (Jane) - will post DAG-Edit user notes
**DONE!! (thanks, Jane!)

Action Item 24 - GO FAQ (Rama and Cath) - populate FAQ with Q & A's

Action Item 25 - Hinxton meeting (Michael Ashburner)
- a: find venue for 10-11 meeting
**DONE

- b: get a Manchester person down to talk about DAML+OIL
**I've asked

Action Item 26a- Hinxton Users meeting (Midori and Karen) - will work out logistics of registration (Consortium members will probably also use the registration page)
**DONE

Action Item 26b- Hinxton Users meeting (Midori) - add suggestion tick box to reg form for what would you like to see
**DONE

Action Item 26c- Hinxton Users meeting (Midori) - mailing to go-friends list asking about desired content/attendance for User's meeting
**DONE (zero replies tho :( )

Action Item 27 - quotes for Virgin Islands meeting proposal (John Richter) - will get quotes and send to list within the next week
**John sent one message and got several replies, so I assume
this is in progress