GO MEETING in Chicago on Oct 13, 14 2001
Northwestern University, Chicago, Illinois USA
Rex Chisholm, host

List of Participants
1. Chris Mungall - BDGP, Berkeley
2. Brad Marshall - BDGP, Berkeley
3. Rex Chisholm - DictyBase, Northwestern
4. Suzi Lewis - BDGP, Berkeley
5. Michael Ashburner - FlyBase, EBI
6. J. Yoon - TAIR, Carnegie
7. Sue Rhee -TAIR, Carnegie
8. Peter Good - NHGRI
9. Trisha Dyck - DictyBase, Northwestern
10. Karen Christie - SGD, Stanford
11. Matt Berriman - Parasitic genomes, Sanger
12. Judith Blake - MGI, Jackson Laboratory
13. David Hill - MGI, Jackson Laboratory
14. Harold Drabkin - MGI, Jackson Laboratory
15. Janan Eppig - MGI, Jackson Laboratory
16. Raymond Lee - WormBase, CalTech
17. Wen Chen - WormBase, CalTech
18. Midori Harris - GO EDITOR, EBI
19. Evelyn Camon - SP-human proteins, EBI
20. Bernard de Bono - visitor from Cambridge
21. John Richter - BDGP, Boulder
22. Erich Schwarz - WormBase, CalTech
23. Jason Stewart - BDGP, Albuquerque
24. Hanqing Xie, Compugen

ACTION ITEMS FROM CHICAGO MEETING - OCTOBER 2001

1. ACTION ITEM: Get temporal and anatomical CVs into CVS from participating databases.

TEMPORAL: need to add temporal ontologies, but maybe we need to add dimensions rather than time or relative temporal terms? discrete temporal terms. like Tyler Stages (Mus; 'life stages ' in worm; template for others?

STATUS OF ANATOMIES

FlyBase done, FlyBase also represents 'derived from' in organ-organ relationships

TAIR.... pretty good, almost done, just checking...

SGD. anatomy...12 terms, done

MGI ... Martin met with folks in Edinburgh, fine for MGI to commit mouse anatomy to the web site. Adult anatomy is not complete, but we could contribute it. Defined by Edinburgh by anatomical space definitions... 3D... Embryological one is done.

Worm... Wen Chen...has been working on this. Life Stages for embryos is done. It has been converted into GO type form. Need to do refinement on definitions. Right now, Wen has built structure with 55 terms.... Worm... can keep temporal aspects. In C. elegans, because of knowledge of every cell, can actually doing anatomy in terms of big picture of all cells. Raymond has done a pilot on the feeding apparatus, the pharynx of worm. Working with David Hall of Einstein to work out anatomy (also Sylvia Martinelli at Sanger has some work to incorporate).

2. ACTION ITEM: need a tool to create cross-product terms

see further discussion of this point under BDGP report

3. ACTION ITEM: Post new biological process that incorporates updating developmental processes. Ask for comments by Dec. 31. After that date, do the update and commit it. Post new Biological Process.

4. ACTION ITEM: come up with initial default GO slim, send around for comment, incorporate into database such that db changes could be flagged...then we could make that available as GO-SLIM. Ask that people who make variants and publish with that will post them...Midori and Michael will create default file... technically...flag in flatfile... *Note, this is a carryover from Bar Harbor meeting*.

don't need 'static' one...???
but do need easy way to create one, need 'this one, that one, all the others'
last time we agreed that we would
a) archive any that are used
b) people will use this feature a lot in the future, so how do we make this easier
c) we decided to provide a default one....

when the database is implemented, this will be easier to manage...
d) people that work with alternative GO-SLIMs will post them

5. ACTION ITEM: Agenda items for next meeting

1. Chris Mungall... intro to ontological formalisms
2. Michael... will report on the literature of relationships in ontologies, WordNet book report

All terms need parents. The downside of using a quick upgrade of everything is that 'is-a' relationships may not be correct for all, particularly biological processes. Should we add an 'is-a' to the top level? Also, the 'part-of' relationship is complicated since we use 'part-of ' in different ways. Need to investigate the implications of this. We may decide that we don't want to get more complicated than we are.

Do it as needed, well, why would we need it...
1) to use outside ontological tools
2) to resolve multiple uses of 'part-of' types
3) to facilitate doing queries

6. ACTION ITEM: standardization of database abbreviations

Michael will finish off flatfile in one day of standard set of linking database abbreviations. He will post to CVS.

done: go/doc/GO.xrf_abbs

7. ACTION ITEM: New Evidence Codes

1) invent an evidence code for 'nothing is known biologically' ND; ND:evidence-reference would be a local database citation that abstracts to methodology.

Done

****

the next one came up during the user's meeting and didn't receive discussion from the full group....may have to wait until next meeting for consensus, or may be agreed do by email...

2) proposal at User's Meeting to add an evidence code for 'inferred from curated orthology'; ICO; evidence-reference would be a database citation that abstracts to the methodology. This depends heavily on a shared understanding of the term 'curated orthology', but in this first instance refers to the case where RGD is transferring GO associations to Rat genes from MGD via the curated orthology relationship provided by MGI. Further discussion reveals the complexity of this. Orthologies are most often defined at this time by sequence similarity. The transference of functional annotation therefore might be considered 'ISS'. EXCEPT that this really depends on the method of assignment of function to the orthologous object to begin with. If the assignment, for example, of a function to a mouse gene was as a result of a biochemical assay, then ISS for the rat protein might be appropriate. If the assignment of function to the mouse gene, however, was via an electronic assertion based on (perhaps) shared domains, then

this would be an IEA assignment in mouse and even with the orthologous relationship to a rat gene, the rat gene GO assignment should not be given an ISS evidence code.

8. ACTION ITEM: We will add a date field to the entire gene association file
We are adding a date column, mandatory for all annotations, YYYYMMDD. It will mean the date on which the association was made; it will not need to be broken down into "created" vs. "updated," because we "update" annotations by adding new lines to the association files (and deleting old lines if the situation calls for it).

The date field can be used in conjunction with the ND code, so that curators can tell when it was that nothing useful was found. This was the original motivation for including the date, but we quickly realized that dating annotations was good for other reasons as well.

This has now been documented in the GO documentation at the Web site.

9. ACTION ITEM: Update Documentation to explain new use of the TaxonID column in the gene association files. Revised syntax for TAXON ID Column is: 1st ID = taxon encoding the gene product; 2nd taxonID refers to the context, i.e. the user organism of the gene product. Syntax will be taxon1 ! (pipe) taxon2

Done

10. ACTION ITEM: Change requirement for submission of sequence information for gene products. Remove sequence subdirectory and replace with subdirectory holding files of gpID: proteinSeqID from the participating db. BDGP will use these files to yank protein sequences and generated appropriate def line according to current GO standards. Resulting sequence sets will be posted regularly. Peptides will be the sequence type.

new syntax: DBID:geneObjectID DBID:seqAcc#; DBID:seqAcc#;DBID: seqAcc#; where multiple seqIDs are only added to reflect alternative transcripts, not allelic variants

11. ACTION ITEM: Update directory structure on GO web site. Karen will transmit this directory restructuring to Mike...

gp2protein
- gp2protein.sgd
- gene2protein.mgi
- gene2protein.etc

/protein2FASTA created by se group, will go into monthly archive,
also get rid of species subdirectories, create new one 'gene-associations' and put all the mod association files in there.

rename 'monthly' dumps to an informative name
Parent directory 'Data Snapshots' subdirectories 'Current' and 'Archived' ??

SO monthly_downloads (cvs tag on the 1st of every month) subdirectory will be
/currrent_yyyymmdd
- ontologies
- xml
- db
- gene-associations
- definitions
- database load
- sequence set

/ archived
under that, the monthly subdirectories moved from the previous monthly

all the other directories will be the most current... the monthly is the snapshot version....
remove 'abstract' directory
remove 'archive' directory
'docs' okay

'external to go' ok
'mail' ok
'note' can be deleted
'ontology' keep
'schema' goes
'sequence' delete
'software' delete
'xml' delete

12. ACTION ITEMS FOR JOHN RICHTER

*need to attach cross-product terms to existing ontologies
*need to be able to track identifiers to components of the cross_product
*will notify users 'there may be dangling references here'
*need to address 'merging' issues.

John will release 2.7 next week, in 2.8 will allows dangling objects
other plug-ins that are being suggested.
gene product fetcher...will get from database
also, just select the ISBN for some select set of references, including at least the Oxford Grid ISBN number.

John is working on an html toolkit that will show java trees on the web...a little servlet.

13. ACTION ITEM: User's guide for DAGedit... John will set up a WIKI page, and anyone can contribute.

14. ACTION ITEM: Transition into using database as primary repository

John will work on history tracking mechanism while we test the db and flatfile saving issue. On a Friday, John will say 'we're going to start using the database'. John will populate the db on the weekend with the newest stuff on CVS. Curators won't do anything over that weekend. From then on, curators will save to db and at the end of the session will also export to flatfiles. This will continues as long as we need to . At some point, John will have us revert to the old system if needed. It's important during that time to check email before using to check for recent messages. John will try to give us notice. Also, won't due before Nov 2, but are tentatively schedule this event (the email msg) for Nov. 2.

ACTION ITEM. Continue to consider the need for a DBA.

ACTION ITEM: Need a UK mirror of the GO site....Midori will talk to Pete.
Check with Chris Richter--he talked with Pete (Petteri Jokinen, EBI systems)

****

<u>ACTION ITEMS FROM BAR HARBOR MEETING - July 2001</u>

1.Go Slim

a). Consensus that there needs to be a new GOSLIM developed. A small working group will select terms for GOSLIM.

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

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

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

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

2. Changes to syntax of gene associations files

There is a need to define the object being annotated explicitly. Changes to be made are: 1) add a column that defines the object being annotated (or the moment the options will be gene, transcript and protein).
2) The symbol used in the association file will be the symbol for that object (e.g. if annotation is to a gene object then symbol = gene symbol, annotation to protein object then symbol= protein symbol). Same holds for synonyms, they should match the object being annotated,
3) add a column for TaxonID (from NCBI) that defines the taxonomic node for the organism whose gene/protein/transcript is being annotated.
4) Midori will update the web pages with the new information/format for the association files. An XML format will be described to export the association files.

3. "is this a function or a protein name?"

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

4. Web site management

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

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

Reports from Consortium Members

1. Rex Chisholm, DictyBase

Warren Kibbe, getting the database set up, making a few GO associations for an interim collection of genes, 5600 full length cDNAs, ~ 70% of genes (between 8 and 10,000 genes). NIH grant pending which would provide some more curators.

2. Suzi Lewis, BDGP

The next thing to accomplish with DAG-edit is to connect to the database, i.e., have the DB directly connected to the editor instead of the flatfiles. Once this is done, a secondary goal is to include plugins and to add a history viewer. Also, we want to switch to the database, and we want to have synchronized monthly updates. Every month (around 1st of month), take snapshot, export XML and to database (don't rewrite back to flatfiles). The XML and DB include gene associations. We have added another format of the data, RDF (an extension of XML) this will get us to DAML-OIL, Protege, etc). Chris has gotten sequence data...fly, yeast, Arabidopsis... BLAST search would bring back hyperlinked output with a small subgraph.... Sequence data is on the GO site only for SGD. Otherwise have to download...

3. Michael Ashburner, FlyBase

Have appointed new curator, Rebecca ,will start Nov 5, FlyBase inherited a large number of electronic annotations. An editor has now looked at them all and there are no more IEAs. Several thousand proteins from Celera had no GO data. Michael has been through all of those and has been able to assign GO terms to about _ of them. Still there are ~380 genes with *GO annotations but no references*. Michael is working on that. Regular FlyBase curators added GO terms as well. Re-annotation of release 3 has just started. Release 3 / Drosophila should be complete with no gaps for euchromatin and should have no ambiguities. Now true for 2 chromosome arms. Harvard and Berkeley trained 10 people to use Apollo to annotate fly sequences...Plan is to have reannotation by April 1. 15 min per gene....

Improving Cross-Links...with PIR, database of modified AA cross links with definition files 2) Minn. Biodegradation 3) MetaCyc...doing just pathways, working with Peter Karp
following ISMB, 4) comments back from MIPS...They are interested in making a MIPS 2GO mapping available, Michael and Midori have both talked with Klaus Meyer.

4. Sue Rhee, TAIR

Manual annotations have started. Tools developed to do this (PubDB). PubDB stores matches to gene names and keywords (GO, Anatomy) to papers. Curators can validate the matches and update/insert new genes, and gene aliases. Currently developing the web forms to validate and update and insert new annotations between genes and GO terms. Will package it and make the source codes available in a couple of months. A lot of this work supervised and carried out by Leonora. Leonore will be leaving TAIR, she is looking for more of an education/outreach kind of job. TAIR is actively looking for replacement for her. J. has been working with Leonore. Rest of curators at Carnegie are working with the GO, learning to do manual annotations. There are about 4,000 annotations using GOFISH methods to 3890 genes. Added 50 hand annotations to GO. Using PubDB matches of known gene names and GO terms within papers to screen literature and to provide a set for curators to work with to annotate genes. They use GO Editor, and other tools. So, they take whole set of GO terms, run them against abstracts, and make a file of matches for curators to manually validate using Web forms. Doing the same for gene names gives about 80% validation rate when examined. A lot of gene matching to papers, over 90% have GO term that match as well. TIGR curators will use PubDB and share the literature curation efforts.

Two types of electronic annotations have been done. Nicky at InterPro has provided InterPro matches to Arabidopsis proteins. TAIR also ran InterproScan and used Nicky's InterPro to GO mapping. Nicky may have used set from MIPs. So, slight differences, but will submit Nicky's annotations to common protein set to GO. There are about 10,000 GO annotations out of about 20,000 proteins. For the remaining 5,000 proteins, will put in TAIR annotations. Currently, they are in the process of removing annotations that don't make sense for plants.

Working with Peter Karp at MetaCyc to add plant pathways to MetaCyc. We used the pathologic script to find pathways matching to Arabidopsis enzymes.. All plant proteins thought to be enzyme or enzyme like (~6000) were passed through and ~1800 proteins were matched. 112 pathways have more than 50% reactions matching and 64 pathways have less than 50% reactions matching. We are waiting for Michael to finish going through MetaCyc to GO mapping and will submit the GO annotations from this after manual check (Lukas Mueller, mueller@acoma.stanford.edu) Lukas Mueller is one of the TAIR curators who is very interested in GO.

Anatomy development. 247 anatomy terms and 54 dev. stage terms. These will be submitted to GO once the new CVS architecture is set up. These have been submitted to the CVS repository for Plant Ontology Consortium Lincoln Stein has set up at CSHL. Will do comparisons to combine to create higher nodes once Gramene submits their ontologies.

They contacted Paradigm to develop collaborations as they provide extensive service for phenotyping for plants.

TIGR will use PubDB to annotate genes to GO. TAIR has provided login for them. We will discuss on the process of operation. We decided to separate labor based on processes. TIGR has microbial systems in GO annotations. TIGR and TAIR cleaning up genome annotations together.

5. Karen Christie, SGD

SGD has finished off the Oxford Dictionary. Marcel Mendoza, summer intern, went through word files from scanning dictionary, moved to RTF. Mike Cherry wrote web interface to query... password accessible. RTF files allow one to know the difference between the term definitions and string. Have added 2000 terms since July and now have about 17% of them defined. Internal progress reports including progress on GO. New curator Rama Balakrishnan, she will do GO annotations

Mike, Rama, Karen went and meet with Russ Altman's group. He is moving into the genetics department and will have more interactions with SGD and GO. One of his grad students is working to put GO into Protege. Text matches between literature and GO terms to develop interface for curators to use to guide annotations. IEAs from Valerie Wood...they are comparing new set with all other IEAs. For genes with no annotations, may be good. Looking at comparing with new set from Valerie Wood.

6. Matt Berriman, Sanger

Matt Berriman, Sanger Institute Just released version 1 of GeneDB (http://www.genedb.org) genome

database for parasitic groups, initially S. pombe, Trypanosome brucei and Leishmania major. Every genome will be annotated to GO, and GO association files will be released as these are done. Have written some parasite-specific GO terms (http://www.sanger.ac.uk/Users/mb4/GO). Annotation of malaria genome still planned to happen before Christmas. TIGR has mentioned they would like to get involved in that too.

7. Harold Drabkin, MGI

Majority of annotations still electronic with over 6000 genes annotated. There are over 1500 hand-annotated genes, and these are done both as new genes are entered into MGI and as curatorial review of gene families and other genes. There has been an increase in hand annotations with function unknown. (*discussed further in the 'annotation discussion')*. MGI curators are focusing on sets of genes and on new genes with no GO annotations. The MGI gene association files are updated on a weekly basis and the new file sent to the GO site and posted on our ftp site. We have modified the associations file to include taxonID, object type, 'non' option and syntaxes as agreed by the group. MGI curators are now adding over 100 annotations per week.

8. Midori Harris, GO Editor

Added a couple hundred GO terms... Most of these have been specific requests from SP annotators. Have identified areas that need work when time allows. Reorganizing documentation. Still need to hire Midori's assistant, but are now interviewing.

9. Evelyn Camon, SWISS-PROT

EBI, Wolfgang Fleischmann is doing automatic annotations to all species. all of GO... of 100,000 SP annotations, 40% have GO assignments... EBI open database to provide access to human gene products...have reached the stage that NCBI/Proteome, SP/InterPro, other GO translations (EC, keywords, etc), SGD, MGI, FlyBase, all annotations all together in dataset...

Oracle DB of all associations... 1.5M entries of (Protein to GO) Will also submit gene associations back to the GO. Each assignment has it's own ProteinToGo ID, can extract information about GoAH. 28,727 eligible human proteins, 11435 IEA proteins, (9636 InterPro true match, 4,000+ via SP keywords. 577 via EC codes, 9662 hand annotated proteins in the human dataset, 6864 done by Proteome Inc.; 2,830 by EBI/SP curation team. There will be a new release of InterPro in the next couple of weeks. 3,000 human entries identified by Paul, those by Proteome were removed from curation set. SP curators have been working hard to assign GO terms to the remaining... This stage of the work is now complete. From now on, SP curators will annotate GO terms. Report from Nicky about how InterPro annotations are done. QuickGO browser has some new functionality. Now there are links to microarray expression database. Proteome analysis pages at EBI urgently need GO-SLIM.

10. Erich Swartz, WormBase

IEA has done 1/3 of 19,000 genes. Creating new parsing of ontologies and expanding automatic annotation. Building an anatomy and developmental timing ontologies. Erich has been trying to finish RNAi ... 52 phenotypic types. Currently limited to not having full GO curator. Ideally by next meeting... Proteome has asked WormBase for help them with 'GO-izing' their standard vocabulary that they use for all the phenotypes. Erich
has been in contact with WIT2 annotations group. have set up a collaboration with them to do that.

11. Physiological Presentation from Bernard de Bono. Bernard has been lecturing on human physiology for the past 6 years. At the MRC-LMB he is annotating protein repertoires from the human genome, and is therefore interested in bridging representations of physiological processes with gene products.

During his talk he suggested a physiology model in which a biological process could be precisely defined in terms of a large-scale exchange between compartments. Four main types of compartment were defined: subcellular, cellular, extracellular and surface. He created fourteen tissue Cell Blocks: seven of them interface with the extracellular compartment only, while the other seven interface with the surface compartment as well. As Cell Blocks are the

terminal leaves of a classification tree, it is not intended that a particular histological cell type should belong to more than one Cell Block. The human Cell Blocks described are:

a) Cutaneous, Respiratory, Urological, Uterine, Testicular, Gastrointestinal, and Placental
b) Hematological, Endothelial, Endocrine, Muscular, Nervous, Skeletal and Connective Tissue

The whole technical discussion covered the following points, points 1. and 2. having been described above. Points 18 to 21 involved suggested extensions to this model to be discussed at a later stage of development.

1. A Process is the exchange between one compartment and another.
2. A series of major functional Cell Blocks was created and classified in terms of compartment contact.
3. An organ then becomes a Cell Block composite.
4. Cellular and subcellular compartments can then be addressed by the location of the Cell Block.
5. Extracellular and Surface compartments can be addressed by the Cell Blocks that are bound by it.
6. As organs are Cell Block composites, the anatomical location of the Cell Blocks can be tracked down.
7. A Process is an objective that can be depicted by a series of sub-objectives that may have to be temporally sorted.
8. A Process may occur only during a specific milestone in the organism's stage of development.
9. Separate time scales represent 7. and 8. to become the temporal ontology.
10. A spatial ontology represents 4., 5. and 6.
11. The Location Ontology captures 9. and 10.
12. A Process then becomes a cross product of Location Ontology and Function Ontology.
13. The Process Ontology editor creates paths in this co-ordinate matrix and assigns Physiological alias to every path. From 7., a path may have subpaths that are sorted along the temporal ontology co-ordinate.
14. The organism database GO annotator generates ontology co-ordinates in terms of What (function ontology), Where (spatial ontology) & When (temporal ontology) for every sequence.
15. If a gene product's ontology annotations hit a path from 13. the gene product automatically inherits that Process.
16. Cross product annotation is space efficient and robust to updates. Process definitions are more precise.
17. The creation of a 'Tool Box' of basic physiological objectives is therefore feasible.
18. Compartments can be mapped into partitions.
19. Processes description can then extend to exchange between one partition and another.
20. An Enzyme can be seen as pulling Molecule A out and pushing Molecule B into the same partition.
21. Can embryological/developmental processes then be defined as a transition from one Cell Block to another?
22. Caveat: this cross product can be represented by a DAG, but needs more than just 'is a' and 'part of' type of edges.

Conclusion: As more and more complex organisms are annotated at gene level, it will become increasingly evident that gene products participate in more physiological processes than practicable to annotate directly - suggesting that curated paths using the spatial, temporal and function Ontologies as co-ordinates may be the solution to represent physiology.

Annotation Issues

1. Midori and Rex continue on search and destroy mission....to get gene products out of ontologies...

2. DEFINITIONS.... now 18% done. Tomorrow we will be looking at the Oxford Dictionary of Biochemistry and Molecular Biology that will be provided as part of the Editor. Keep Oxford definition and add local identifier... For edited references that come out of some other resource, add original resource and the modifier resource. So, with more than one citation, it means that definition is composite of the two resources. Good progress. ... including as it does new terms that are supposed to be only entered with a definition....

3. Database abbreviations are not always consistent. We will make a little flatfile of these... database:identifier and will post to CVS....

4. Annotation Discussions

1) Annotating to 'unknown' is different than annotating to 'didn't look, don't know'. So, when annotating to 'unknown' because biology isn't known, **MGI** puts 'unknown' and references the paper... go through all the papers that are for this gene, use the most recent paper as a reference...last paper about that gene that was looked at. **SGD**..if they look and there are papers available but they don't address the issue, than they use a generic SGD citation...so, like MGI, from modification date. **TAIR** associates to the last paper, 'NAS' give the last dated paper. Unknown is used as annotation tool to help the annotation pipeline so that you know the effort was made to annotate the GO. We considered advantages and disadvantages of both approaches (MGI & SGD) before coming up with the ND solution.

SGD: cite generic SGD

Advantage: doesn't attribute statement to a paper that didn't actually contain it.

Disadvantage: no indication of when someone last looked for information.

MGI: cite most recent papter

Advantage: provides a date so that curators know to look only at more recent papers for additional information.

Disadvantage: implies that the paper actually stated that something was unknown.

Using ND in conjunction with the date added to all annotations captures the advantages of both previous approachers.

So question arises, what do we put on the GO site?

Have ontological term.

1) invent an evidence code for 'nothing is known biologically' ND
2) evidence-reference would be a local database citation

2) ALSO will add a date field to the entire table...so will know the last annotation date...for each line in the gene association file....to the end...yyyymmdd

This will mean that the date on which the association was made; it will not need to be broken down into 'created' or 'modified' because we 'update' annotations by adding new lines to the association files (and deleting old lines if the situation calls for it).

VIRUS/PATHOGEN/PARASITE

Virus using host gene products will have associations to the virus genome. Should use the gene reference to the model organism database. Issue is how to annotate, for example, mouse protein that is abnormally functioning in the normal of the viral genome. That is to say, the function of the mouse protein is 'normal' for the viral process, but abnormal for the mouse process. We need a way to identify the genome of the process being annotated. There was intense discussion reflecting that a curator of mouse proteins wouldn't be annotating to the viral process because that process is 'abnormal' for mouse biology. But the curator of viral proteins might want to indicate that a mouse protein was 'part of' the viral transcriptome, or something like that. So, it was concluded that in that kind of instance, the curator was annotating a normal process, and the association file needed to indicate both the taxon of origin of the gene product (which is the function of the taxon ID now), and additionally, be able to indicate the taxon of the genome being annotated. So, if only one taxonID is presented, it means that is the taxon of origin for the gene product and the taxon of the genome being annotated. If two taxon IDs are presented, then the 1st one is the taxon ID of origin for the gene product and the second one is the taxon ID of the genome under annotation.

The important point here is that what 'normal' means is relative to the organism being annotated, *i.e.,* normal for the host vs. normal for the virus.

****

CONCLUSION, TAXON ID Column, 1st ID = taxon encoding the gene product; 2nd taxonID refers to the context, i.e. the user organism of the gene product. Syntax will be taxon1 ! (pipe) taxon2

Implicit here is that the user taxon indicates whose perspective of 'normal' applies.
****
Columns in SP annotation files
db contributing - identifier from the contributing database (SP, TrEMBL, international protein index) - 3rd column would be international protein identifier -
from other sources, 3rd column would be the gene symbol...would put any gene names in the synonym field...wanted to be able to use the IPI... so our suggestion is that they use the gene name if they have it, use the IPI as a default, if they update gene name, the IPI moves to the synonym....
****

BDGP report on software and database development for the GO

John Richter, software...DAG-Editor...version 1.207,

1) can load files directly from the GO site,
2) new plugins load automatically, but can disable that feature
3) can add new relationships

Chris Mungall...GO database

1) monthly archives in database, XML and flatfiles
2) have been expanding schema....can now have sequences in database.
3) want not the FASTA, just the relevant protein seqID, these will be loaded into the GO database. We do have sequence directory on CVS, which has SGD file of sequences on it. This will be dumped and new subdirectory created as noted below.
4) building tools to help in the proper use of the files by the community. One example would be to have triggers to prevent others from grabbing files and using them in analysis without understanding the IEA or levels and evidence.

Cross products

make a new ontology (transcription occurs in the nucleus)

move nucleus to this new ontology
move transcription to this new ontology
'new term' nuclear transcription, has new relationship

should be high priority on the list of things to do to be able to create cross products. Will put the cross products where they belong. So, do we create a huge file of all, or do you create cross products with dangle unfound terms. Why don't we just permit loading of the entire directory....also would need definitions. So the load would be huge, and what would be the point. If we were loading 10 different files all the time, why not load as one big file. Mode of operation is to allow cross product generation to users...so the real question is if a new process term, a cross-product term, were in the process file, there should be access to all the term components of the term including the anatomy terms...Have to support dependencies in the different files. So if you load process ontology, will be prompted to load the anatomy files. 'verified' in the sense that a curator has looked at it.

GO DATABASE Chris Mungall

sequence blast results multiple sequence viewer... width of sequence bar reflects the degree of similarity. shows the multiple sequences on the top and then the blast results underneath. GADFLY
So question now is do we move into database? As in, do the curators edit into the database rather than into the flatfiles, and at the end of each day, commit as well to the flatfile. So, need a group to manage this database, someone familiar with MySQL, postINGRES....

BROWSERS - BRAD MARSHALL

AMIGO...www.godatabase.org very soon....(on the internal LBL site)
Future directions

**** BLAST server so that you can do a BLAST search against the gene product sequences with links back to the tree
****get a portion of the tree that you like, and get a FASTA dump of all the sequences associated.
****new browser gives number of terms annotated to term or terms under it.
***coming soon, will be able to select terms and download into a FASTA file
****want to select more than one evidence code

****want to filter by NOT for one or more evidence codes

****curator approved is everything other than IEA

****advanced search, can search by gene products, or gene symbols, pick data sources, etc. used to have ability to paste in a list, but have taken that feature out. but now people are again asking for that. So, can put back that capability...

****have extensive docs for this...

Jason Stewart, new to the group, may do some software development with GO. He is also familiar with MGED development of structured vocabularies. MGED has been working with Rosetta, Affymetrix, others, to have a data model. microarray array gene expression. MAGEML is the mark-up language... just had a jamboree in Toronto and put all the source code in Source Forge. They are building annotation tools to help build the XML files. So, the model is very large, has 146 different classes, very interconnected, a lot of context. There are lots of parts of the model where 'terms' need to exist. Simple and complex, some are forms of restricted vocabularies, some of them will be real ontologies. What they did indicate in the model that whenever they come on one of these terms, they designate as an 'ontology', e.g. this is Jason's ontology for describing spots on a glass file SO when others use the file, and find an ontology term that they don't know about, they will have a url to find the ontology. So the program that is taking the XML and putting in the local database needs to go get that ontology and put it in the database that is being developed. SO Jason is writing a perl script that will allow researcher to go get the XML at the url link to load into their database. So, researchers can each submit their own ontologies... So everyone around the world can utilize the terms now...*not that this will not facilitate the development of a community standard*

MIRROR...need a UK mirror, Midori will talk with Pete

VERSIONS, REVISIONS, RELEASES... Diff problems are due to genuine bugs in DAGedit that will be fixed in 2.7. Bo at AstroZeneca asked if we could include a diff along with the monthly. No, we decided not to do that, the new directory structure should help this issue, since all the files will be organized better. Term Counts...Bo...Mike and Midori got the same numbers...no one knows why BO got different numbers.

GOBO - global open biology ontologies

We, the GO consortium, have three common GO ontologies. Other ontologies such as anatomy and temporal ontologies will be developed and 'owned' by MOD databases. At the Bar Harbor meeting last July we heard in particular from plant people about the need for phenotypes ontologies. Also need an ontology for biological substances so that those terms can be taken out of the function ontology. SP crew may do this. So, we need an umbrella under which we can have a variety of ontologies. This would essentially be a web site, cvs, or ftp site onto which different communities would be encouraged to deposit their ontologies.

1) The ontologies would be open.
2) They should be instantiated in GO syntax, flatfiles, so that they could be used with GO tools.
3) They should be orthogonal with existing ontologies...this is the hardest to resolve...
4) They would share ID space
5) Definition files should accompany ontologies.

Orthogonal issue is the biggest concern. There are reasons for this. For example, there are alternative, competing ontologies in the same domain; by definition they are not orthogonal. We would want to distinguish between competing and complementary. We would need to explain why orthogonalities are there if they are.

So, would be fine if GOBO was a web server site...community would be offered to send us their urls...There are other web sites for biological ontologies, but they won't adhere to the five principles above. Also, the anatomies and other ones that are used as part of the GO project would be handled somewhat differently. Most of these, however, will have some aspect of involvement with the GO project.
Suzi will be writing adaptors for RDF / DAML-OIL, a data adaptor for GOedit. The beauty of using DAGedit is that you get something that can be instantiated in GO syntaxes. **ACTION ITEM...Michael proposing to publish a short editorial about this.** ...

GO COMPLIANCE and JOURNALS.

Michael reported that Nature has been doing the experiment of using GO terms as metadata to articles. Michael had a discussion to Declan Butler about this. This might be a more interesting concept than 'compliance'. Nature editors would do this. Coming down to 'keyword' for the article that is selected to from the GO. Rex says maybe if up and running with Nature, we should suggest the concept in a letter to other mainstream journals. GOSlim could be a keyword list, but for article keywords, should be any GO term. Evelyn says...keywords...EMBL adds keywords to flatfiles. Curators at EMBL talk to 8,000 scientists a month. Authors add keywords; they could be encouraged to add GO terms as the keywords. GenBank might add as a dbXREF. dbXREF ... GO has not been accepted as a dbXREF... but Michael is going to next advisory meeting . GOdbXREFs with SP? Midori says they're going to be stored in SP-Oracle DB and will be visible to the public sometime. **ACTION ITEM**... Michael will try to get GO accepted as a dbXREF for the sequence databases

JOBS and FUNDING

Midori's assistant advertised
FlyBase curator of GO hired (Rebecca)
Evelyn hired
Lincoln's grant is going in.

GO jobs will be on a separate Web page

Judy will do Grant Progress Report due Nov 1. This year TAIR and WormBase will get funding from this grant.

PUBLICATIONS

Michael is talking about GO at Novatis meeting in London that requires a publication Michael will do
Matt has TIG paper to add to the progress report
David and Michael will continue on cross-product paper.
Panther/GO comparative paper...evaluation on electronic annotation...David and Suzi (FANTOM)
Cathy Ball, SGD - 4way comparative paper...did pull in all GO annotations that were available...still working on publication...
Matt has forthcoming book chapter with Midori to go Current Protocols in Parasite Genomics
Han Xie...internal review of Compugen stuff...
Also…
Berriman, M., Aslett, M., Hall, N., Ivens, A (2001) Parasites are GO. Trend in Parasitology 17(10) 463-4. PMID 11642257

WEB PAGES

Midori as working on various pages
Karen and Midori talking about totally redoing Web page

****Separate page for the job listings

****Page for getting in touch with GO, listing email lists and other contacts, participating groups would list particular contact for that group

****Another page of all members and former members

time on this? Karen will send out url on personal space before posting publicly

NEXT MEETINGS

GO Users Meeting Feb1, GO meeting on Feb 2 and 3rd (O'Reilly meeting the 3 days before them).
Academic staff get 25% discount...$600, Faculty get 50% $495. for O'Reilly meeting
How many people from us will be there...? 25...
Might be advertised more heavily...how many would be expected at Users Mtg? over 100 at least
Would have to pay for network access. **yes it's critical we decided.**
Next beyond that...Michael with host in Hinxton....
Next beyond that...maybe hosted by Compugen in Princeton...
TIGR - contact Michelle and ask for a TIGR rep