Meeting of the Gene Ontology Consortium
Cold Spring Harbor Laboratory, Plimpton Room
May 12-13, 2002

Sunday May 12th, 2002e

ATTENDING:

- Suzanna Lewis	FlyBase	Berkeley, CA
- Michael Ashburner	FlyBase	Cambridge, UK
- Karen Christie	SGD	Stanford, CA
- Judith Blake	MGI	Bar Harbor, ME
- Elizabeth Nickerson	GK	CSHL, NY
- Janan Eppig	MGI	Bar Harbor, ME
- Courtland E. Yockey	AstraZeneca	Delaware
- Matt Berriman	PSU(Sanger)	Cambridge, UK
- Katya Mantrova	Incyte Genomics	Beverly, MA
- Han Xie	Compugen	Jamesbrook, NJ
- Liat Mintz	Compugen	Jamesbrook, NJ
- Bernard de Bono	MRC	Cambridge, UK
- Michelle Gwinn	TIGR	Rockville, MD
- Linda Hannick	TIGR	Rockville, MD
- Harold Drabkin	MGI	Bar Harbor, ME
- David Hill	MGI	Bar Harbor, ME
- Rex Chisholm	DictyBase	Northwestern
- John Richter	BDGP	Berkeley, CA
- Pankaj Jaiswal	Gramene	Cornell, NY
- Susan McCouch	Gramene	Cornell, NY
- Martin Ringwald	MGI	Bar Harbor, ME
- Midori Harris	EBI	Hinxton, UK
- Eurie Hong	SGD	Stanford, CA
- Chandra Theesfeld	SGD	Stanford, CA
- Mike Cherry	SGD	Stanford, CA
- Doreen Ware	Gramene	Cornell, NY
- Chris Mungall	BDGP	Berkeley, CA
- Eimear Kenny	WB	Caltech, CA
- Lukas Mueller	TAIR	Carnegie Inst., Stanford, CA
- Daniel Barrell	EBI	Hinxton, UK
- Evelyn Camon	EBI	Hinxton, UK
- Becky Foulger	FlyBase	Cambridge, UK
- Jane Lomax	EBI	Hinxton, UK
- Amelia Ireland	EBI	Hinxton, UK

GROUP REPORTS

Database summary - Suzanna Lewis

- presented table of all non-IEA annotations
- number of terms has increased 10% since last meeting (now 22,000 F; 21,000 P; 15,000 C)
- idea to show % genome covered, buts gets into issue of estimating numbers of genes
- new ec2go mapping (thanks to Daniel)
- pie charts on Amigo!!! can break down by group, by GO-slim like terms (chosen by numerical representation)
- Michael Ashburner suggests being able to choose a specific GO-slim file
- Matt suggested being able to keep top level pie chart when going to breakdown of a slice

FlyBase – Becky Foulger

- Fritz Roth data set added

SGD – Karen Christie, Mike Cherry

- EC definitions (~1500), curators nearly finished checking these; Rama will add when done
- added ~300 Component annotations from a large scale analysis paper from Michael Snyder's lab (Kumar et al. 2002); added only when 2 different methods confirmed the localization
- working on some new GO tools to incorporate into SGD: one maps sets of genes to GO-slim terms – still in the prototype stage
- for annotations to the unknown terms, have changed over to use of the ND evidence code and have added the date column to our gene associations file
- have changed our software to be able to display NOT annotations
- 2 new curators, Eurie Hong and Chandra Theesfeld, both in attendance

Action Item 1a (Brad Marshall): make display of NOT data possible/correct in AmiGO (e.g. FBP26 for SGD; FlyBase, others have more)

SwissProt – Evelyn Camon

- GOA file re-released – human data only
- annotation for other species – 2.1 million associations , viewable in QuickGO
- next priority – organisms not covered by a MOD
- new SwissProt keyword file, on the website (74% of keywords mapped to GO)

PSU (Sanger) – Matt Berriman

- another organism for *Plasmodium falciparum*, 74% of genome annotated ~3000 annotations, all non IEA
- tsetse: working on gene association file soon, based on BLAST hits, etc,; sequencing is done 21000 ESTs, ~8500 seqs
- life cycle stage ontology
- http://www.sanger.ac.uk/Users/mb4/PLO/

MGI – Harold Drabkin

- GO now incorporated into the MGI database, so MGI browser now faster than it was, not using flat files, and reading data current to within a day, allows Boolean operators
- Martin just put mouse anatomy file onto GOBO
- David has been working on a scheme which could be used for GO-Slim (text of document distributed at meeting attached to the end of this file)
- phenotype ontology – making it loadable into DAG-Edit

TAIR - Lukas Mueller

- added a GO annotation search to the website
- literature curation tool (available via GMOD), have used the tool at TAIR, annotations haven't yet gotten to web site
- added/rearranged terms in metabolism for plant specific pathways,
- embryogenesis vs. morphogenesis, in plants morphogenesis is not an obligatory child of embryogenesis, Tanya's proposal for revisions should be available for discussion soon
- GO-slim version, plant version for TAIR

TIGR - Michelle Gwinn (Comprehensive Microbial Resource)

- CMR: many associations, but can't release them until the genomes have been published, one paper has been submitted, as associations become available will be added to the GO annotations table
- schoenella
- *Bacillus anthracis*
- klebsiella
- auto-annotation tool to assign IEA annotations to non-TIGR genomes, may not work until more genomes are manually curated

- Linda Hannick (Arabidopsis)

- 2 new people, team of 5 now
- 20% of Arabidopsis genome is done, approaching by paralogous families of genes, tool to display similarities, stuff, speeds up annotation, everyone specializes in one area, rather than random
- trying to coordinate with TAIR people to avoid duplication of effort
- next genome – *Trypanosoma brucei*

Pankaj Jaiswal - Gramene

- 9000 annotations for rice, mostly IEA
- ontology browser on Gramene
- putting in rice anatomy and temporal files
- trait ontologies in rice, refining structures
- working with MaizeDB to develop resources for anatomy and trait, phenotype ontologies
- working with Michael Ashburner on chemical ontologies

WormBase - Eimear Kenny

- goal to have detailed descriptions for genes by mid-2003
- Andre P.
- Erich working on more extensive gene descriptions
- 2 new WB curators, 1 is in the process of moving to CA, already doing lit curation
- now 3 WB curators working on GO
- WB is developing ontologies: due for release soon
    - cell lineage ontology (Raymond Lee)
    - developmental ontology (Wen Chen) life stages

DictyBase - Rex Chisholm

- EST collection from *Dictyostelium* , 2000+ IEA annotations
- chromosome II about to be completed, mostly IEA annotations
- still working on final schema for database, using a prototype yet
- ontology: anatomy, life cycle, have passed on to David, as simple test of crossproduct
- funding – everything looks good for DictyBase's funding to be approved

GO - Midori Harris

- got SourceForge suggestion tracking running, and is working well, also helps people making request see that there is a line for requests
- Jane and Amelia making lots of progress
- definitions at 30% now!!
- new curators: Amelia Ireland just started, and Cath Brooksbank about to start

Compugen - Liat Mintz

- no non-IEA annotations
- continuing to work on gene associations, including annotations in different products
- just published their paper in Genome Research
- oligo-libraries arranged based on GO terms, genes that are not annotated are often low-expressers
- over 10% of human genome is transcribed from both strands
- hope to release a new version soon

Incyte – Katya Mantrova

- complete translation of protein properties into GO terms
- all databases now annotated with GO terms
- new term suggestions – 90% are getting accepted
- BioKnowledge library by subscription only from June 1st, free trial until then

AstraZeneca – Courtland Yockey (post-meeting addition)

- Incorporation of GO into AZ Bioinformatics Infrastructure
  - Global "protein annotation pipeline" currently utilizes EBI's GOA and NCBI's Proteome annotations as primary public source material
  - All derivative global molecular class databases being constructed utilize GO annotations
  - A global target decision support system (under construction) will utilize GO annotations as well
  - A number of internal groups continue to either use or inquire about GO annotations from the standpoints of microarray data analysis and text mining applications
- Additions to GO Infrastructure in AstraZeneca
  - MySQL database mirror of GO MySQL database set up which includes both public and AZ internal GO annotations, and which will serve as reference set for derivative uses and views
  - Plans to port GO from MySQL to Oracle were not pursued in favor of a MySQL-only solution
  - Internal GO Annotation effort (GOAC Project) now spans >2000 genes and >13,000 annotations

ACTIONS ON PREVIOUS ACTION ITEMS

- SourceForge suggestion tracker – working well
- John Garavelli visiting EBI – helping with terms that have RESIDs
- John Richter will come visit people if they ask nicely (for help, analysis of system-specific oddities/bugs with DAG-Edit/GOET)
- Not done: Chris – have linkouts to sequence in cases such as ISS with ________)
    - ISS with SP:nnnnnn, click on ISS, get new report page
        - with Literature: PUBMED;
        - ISS with SWP:nnnnnn
- GO.xrf_abbs file: stable way for database cross-reference (Ask Brad) – in progress
    - need to invent a metareference for linking for curator refs
    - need to talk about specific columns, e.g. gp2protein
    - clarify abbreviations
- SO document, Michael Ashburner has submitted, people are commenting

Action Item 1b – 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)

Action Item 1c - AmiGO (BDGP): linkouts in AmiGO to sequence in cases such as ISS with ________)

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,)

CONTENT ISSUES

- Ligand, everyone has agreed upon a solution and Jane is about to implement it

- E&M (Embryogenesis and Morphogenesis, not Electricity and Magnetism, this is biology! ;) – Tanya sent draft out Friday, about ready to implement changes to accommodate fact that in plants is separable from morphogenesis

- have not been terribly consistent about "biosynthesis of ___" when we are talking about modifying a residue within a protein; some of these activities are grouped under 'biosynthesis' while others are under 'protein modification' [NB: not the exact term text strings]
    - so for modification of bases or aa residues within the context of the RNA or protein, then it will be only modification, biosynthesis applies when the substance is made as a free substance

  - **post-meeting addition (MAH)**: the case of selenocysteine, which is produced by modification of a serine residue attached to a tRNA. I think the 'not free so not biosynthesis' reasoning applies here too.

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

- proliferation of sensu terms, (Suzi) do we have rules? should only happen in the case of homonym terms, same text strings with unique meanings for each organism

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

- Kirill – don't have a term for 'group transfer', instances of certainly are covered, won't do this (insert a high-level grouping term) yet

- use of 'AND' in a term,
  - Suzi is against it, because there is high probability of violations of the true path, should work to use 'AND' as a grouping mechanism and attempt to use the structure to represent the grouping;
  - 'and/or' is acceptable ??? – will examine these on a case by case basis and see where they are appropriate

- annotation to two different terms with an OR [**NB**: this would be two lines in a gene_association.yfo file with an 'OR'],
  - DAML+OIL has a way of indicating disjunctions
  - Chris and John are not in agreement on how to deal with this, will hash out the options for software solutions to this issue and report back
  - following the above "conclusion" there were some additional group comments suggesting that the better way to approach this is to construct the ontology to have the appropriate grouping terms so as to avoid the need to have an 'OR' join between two associations

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
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 n Item 4a** 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

- Integrity checks
  - o Do we have any rules for integrity checking? are we at a stage where we could?
  - o lets look for:
    - ▪ child terms lacking parentage that they should have
    - ▪ redundant relationships – still some of these

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, i.e. for items to consider for automated integrity checking and submit them to the SourceForge page that John will create

NATIONAL LIBRARY OF MEDICINE (NLM) AND GO (Judy Blake)

- she and Michael Ashburner will be working with NLM to bring GO into NLM and MESH
- has some papers on the topic, "Lexical properties of the Gene Ontology", but in order to map GO terms to NLM terms (of any type), NLM requires definitions for all (GO) terms, when NLM brings in a new system, they are looking to incorporate the new system as synonyms to existing terms OR make new terms if no syn exists
- Michael Ashburner had meeting with Stuart Nelson (head of MESH) and Betsy… (head of NLM), to establish seriousness of GO on this project
- Courtland ? incorporate GO into MESH ,or have a new UMLS ? A: both, looks like very good progress

GO-SLIMS

- TAIR and Amelia are working on some generic GO-slims, one for plants and one for animals, will be very similar, except for some things like no photosynthesis in animals,
- have archived GO-slim which was used for Celera drosophila, will archive other GO-slims that have been used and which can be found, have written a document on GO-slims to be updated with a caveat about obsolete terms in archived slims
- David has a proposal **(see attachment)**, with an example about how to get all membrane things, need to join membrane of cell fraction with membrane of cell, David chose the GO-slim from the DAG and selected bins as biologist , rather than using a computational method to divide the annotations

- Michael Ashburner is against having 'Other' terms in GO-slims, David's GO-Slim highlights some grouping terms that may be missing from the GO
- there was general agreement that the display software of our dreams would be able to generate an 'other' category (on the fly, maybe?) for pie-chart purposes
- we definitely decided not to add 'other' grouping terms into the ontologies
- handling redundancy, when a term may be annotated to two terms, with different granularity, issues about collapsing redundancy
- each GO-slim should have a document attached to it that explains it rules
- ? from Courtland, about being able to use a GO-slim to map an annotation set, Chris suggested that this should be a script, would be nice to incorporate these scripts into AmiGO so that people can use the various GO-slims and use the one of your choice to map the association file(s) of your choice
- Evelyn ? – naming convention for GO-slims
- no problem to have as many as are needed/used, but we will put them into repository

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)

Action Item 1d –AmiGO (BDGP): Incorporate GO-Slim scripts into AmiGO

<u>DAG-EDIT ISSUES</u>

- John will come visit you to talk about, help set-up DAG-Edit if you ask him nicely
- upcoming change of field in DAG-Edit, where the ID will not automatically be GOID, could DAG-Edit will read ID prefix from root term –Action item for John
- spell checking is not done
- integrity checking – not started, no info to do it, will need to discuss what the rules are
- database
  - occasional problems, still complicated
- capture semantics of transport? email from Chris Mungall to Midori…
  - does this mean the thing itself moving, or
  - Chris has a little thing he can display to talk about this
- relationship type choosing is now allowed – John proposed:

- determine new relationship types
- inform everyone of what it is and symbol to be used
- then implement
- – would have to modify true path rule

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

- sure wish we could do cross-products
  - John will do a 'macro' for this
  - Chris proposed being able to select a term in each ontology and have a table generated, where one could select rectangular blocks, David wanted to be able to see 'part of' relationships…
  - John "but that'll be huge…."
  - David "Embrace the Explosion."

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)
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

- How do we handle IDs when we split terms?
  - Currently, the old term and ID becomes obsolete, and both new terms get new IDs, with obsolete ID as a synonym to each new term

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
Action Item 1e - display comment field in AmiGO (Brad) – display comment field in AmiGO

BERNARD'S PRESENTATION : GOAL (GO Active Language)

- Progress since Chicago

- representation of physiology as a xproduct of anatomy, and multiple GO aspects
- it is possible for terms to inherit processes from parent terms
- activity – any GO Function (F) or Process (P)
- compartment (CPR) – can link P and C terms when we know where a process occurs
- A biological process can be formally described as a relationship between two **CPR**s using an **activity**.
- CPRs – is a region of biological space that can be unequivocally addressed using a combination of nodes from the C, cellular, and anatomical ontologies

- *Bolus discoideum* (Latin for disk-shaped round lump) will be the hypothetical model organism
    - o 3 developmental milestones
    - o 7 cell types

- system models processes
    - o exchange
    - o stage
    - o complexing

Bernard's document is downloadable from the MRC-LMB ftp site; he's also got a power point doc there:

ftp://ftp.mrc-lmb.cam.ac.uk/pub/bdb/GOAL_Framework.pdf
ftp://ftp.mrc-lmb.cam.ac.uk/pub/bdb/GOAL_Presentation.ppt

GOET editor (for GOAL)

- John demos new software, which will probably replace DAG-Edit
- this new program is going to use a DAML+OIL like format, which will allow John to make many things that we've wanted to do be possible much better
    - o e.g. history saves, undo, simpler modules for changing a dbxref (currently programmatically difficult in DAG-Edit)
    - o will allow editing of new types of data much easier
    - o DAML+OIL will have advantages for some of the new data types, e.g. SO; also allows some intrinsic restrictions/rules for a given class
- this is available through GMOD project, available on SourceForge

ANNOTATION ISSUES

Concurrent assignments - Evelyn Camon

(Correlations between terms often used together)

- system in QuickGO which gives curator hints which terms usually turn up together, shows up in QuickGO, also will throw up exceptions "weirdness detector" for annotators

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 1f - AmiGO (Brad) – show concurrent assignments in AmiGO

Evidence Codes

- two concerns
- issue 1 – evidence codes for annotations
  - categorization (currently), but does not imply confidence level,
  - in discussions between this meeting and the previous one in Tucson, and from the surveys done for Fritz Roth, it has become apparent that the evidence codes in use now do not provide an indication of confidence, that curators felt that they could not make judgements on experiment quality from evidence code alone
- issue 2 – judgements of sequence believability
  - another practical ?, qualitatively, to develop a system that provides meaning to people running algorithms, which sequences do you believe in? what are rules/criteria for deciding which sequences to use, and also that the annotations are believable

- want a good test set for which to test algorithms, that only contains the genes and those annotations which are deemed to be of good quality
- Evelyn suggested using the QuickGO algorithm for correlated annotation with all the current IEA stripped out and compare to same algorithm run with the IEAs, i.e. does it still predict the same correlations
- consensus opinion to not include IEA or ISS in training sets
- attempt to evaluate training set (no IEA or ISS) quality
- FlyBase Panther calculations – transitive errors, only 4% (ISS)
  - other error type = F errors (F#%*-up, also about 4%)
- to get the clustered set:
  - do within the group
  - use EBI clustering (TRIBE, InterPro, or SP clustering)
  - Liat/Compugen may be able to do the clustering
- use the training set to help develop a tool that helps with annotation

Action Item17a – 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 Item17b – very cool annotation tool (????, highly dependent on above) – use this to develop an annotation tool that utilizes homology clustering

Annotation Tools:

- Talisman can be downloaded from EBI, semi documented, a curation interface for GO in SwissProt, some discussion of transmitting annotations when appropriate to another MOD, currently no programmers to support this tool/program
- Lucas's tool (now on GMOD)
    - searches PUBMED, creates linking table for specified info
    - preindexed papers against GO terms (perl module for text string matching),
    - implemented in Java servelet
    - right now only abstracts indexed
    - Sue Rhee recently found new software for PDF to text
    - mysql database, trying to make it more generic, submitting a GMOD grant to expand applicability

Action Item18 – 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

Consistent term use:

- Midori raised an issue about attempting to make sure that we use terms in consistent ways, Lisa Matthews has offered to send some notes about term use at Incyte

Monday May 13th, 2002

GOBO and SO

SO

- Mike Cherry did some reorganization of the directory structure and put the GOBO stuff into the CVS repository, see http://www.geneontology.org/doc/gobo.html
- Martin put up first bit of mouse on Friday
- SO attempts to provide a controlled vocabulary for sequence features, and types of genes, e.g. whether primary transcript is edited or not, located sequence features and clones and ways to locate them on the sequence
- Michael Ashburner and Suzi will write a supplemental NIH-grant off the GO grant to get a software person to do SO, since this will require DAML+OIL type slots to adequately describe the information types
- Lincoln Stein, Owen White, Ewan Birney, test project , servers to provide data in the same way DAS server using the SO terminology, will help refine the quality of the SO
- John and Chris made some comments about conversion from GO format to DAML+OIL, John suggested that it might be easier to dev in DAML+OIL from the early stages

Biochemical Ontology

- Pankaj has been working on this
- Michael Ashburner has restructured some of this and hopes to release it upon his return to UK next week
- most restructurings to split out classifications by compound type and classifications by action
- also removed 'compound names'
- Pankaj will parse in CAS #s
- will help to do metabolism by cross-products
- MESH is semantically mixed, and heavily biased by pharmaceutical compounds
- CAS is not open to the public
- about 1400 terms now

Disease ontologies

- Rat people are very interested in these, DictyBase as well
- UMLS has tied together a lot of this type of information, though there are some major issues with licensing and public access; but there are many classifications already so do some of these provide a good starting point (JB)
- Michael Ashburner is unaware of anyone doing this as yet
- where does NCI fit into this? doesn't seem to be much of a relationship…
- SMD/MGED may already have some starts on this type of ontology
- also need to make sure that we get definitions into this
- with respect to tying information to human disease (key to much of our various groups funding), it is key that we make the relationships btw genes/phenotypes and relationships to human diseases
- some overlap with phenotype ontologies and Bernard's attempts to describe physiology
- can't easily use SnoMed, restrictions on its use

Cell type ontologies

- Martin wants to do a mammalian one
- already one for Drosophila

Phenotype ontologies

- Michael Ashburner trying to get together a group for this, Gramene very interested

<u>GKB – Elizabeth Nickerson</u>

- integrative database for human biology
    - biological processes
    - biological pathways
    - collaboration with GO, EBI
- top-down approach, from topics down, rather than gene by gene
- ? how to store as a knowledge base, rather than as a database… want to see when the output of one assertion is the input of one assertion
- tried lots of grammatical tagging, outputs were often unsatisfactory
- now: input and output tagging, and linking to GO terms
- and still want to link to references for every assertion
- using Protégé for structuring the data
- but lack a good interface for authors to input data, currently using Excel spreadsheet for authors so that each sentence is associated with metadata in a way that can be imported into Protégé

GMOD: Generic Model Organism Database (Mike Cherry)

- www.gmod.org
- organization headed by Lincoln Stein
- idea is to create modules, small components that can be used
- more robust, shareable, documented software modules
- so that a new database starting up doesn't have to start de novo with writing their own software, GMOD proposals submission close in a week
- what would a new database have to create within their first 6 months in order to get up and running
- initially asking everyone to make everything Open Source, for GMOD purposes, it is defined by definition on SourceForge site
- GMOD site is an open repository of tools that are being made available
- 4 older MODs are to be given supplemental funds for GMOD efforts, with the understanding that if one group is developing a tool, they enquire of the others, how would you use this tool and make it useable for all the groups

Data Distribution (Chris Mungall)

AmiGO

- pie charts
  - Matt's already suggested a slight modification (keeping original pie)
- Graph view (from term pages)
  - some modifications to clarify (GOID #s)
  - call for suggestions for using the network graphs

Action Item 1 continued - AmiGO (Brad Marshall) – additions to AmiGO

- add a SourceForge site for AmiGO bugs/requests
- gray out obsolete terms (post meeting addition)
- link from treeview page to graph view
- search function for the comments
- don't automatically toggle to gene product when the search result comes up null
- need to make sure that definition references go up with the def, not in the general dbxrefs
- add ability to upload files for multigene search
- GOST, request for it to accept a seqID
- want to be able to search with SwissProt accession numbers (this requires a gp2protein file for every organism, nothing for TIGR, PomBase,)
- having a way of hiding/deselecting GO terms in BLAST report that you don't believe

Chris has some experience with dividing TrEMBL into reliable and non-reliable, may be helpful to others in generating gp2protein files

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

Monthly Releases

- request for synchrony between flat file releases, Definitions at the same time as ontology files
- ftp site is being updated hourly (15 after the hour)
- Courtland Yockey – Could we use the archives to track our understanding of biology
- Courtland Yockey – suggested a month-to-month diff file
  - is interested in this for corporate/pipeline people for being able to track and find differences, and that GO could provide this a resource for others
- Courtland Yockey suggesting some sort of monthly summary of major changes in a place where it is easy to find for part-time users of GO, monthly release notes? this could also explain motivation for changes and clarify rationale

Action Item 20 – monthly release notes (GO) - take a look at doing monthly release notes,

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

Database beta test is over for now

- hopefully will resume at next meeting
- with the new GOET tool

Planning Ahead;

Upcoming Meetings,

- Genome Informatics – John, talk about GOET and relation to databases
- ISMB – Michael Ashburner to give plenary
- November meeting (17-20) in Hinxton – MGED meets GO, w/ diseases and chemicals
  - ontologies, and tools for building ontologies
  - Suzi soliciting Bernard to submit abstracts to this (Michael Ashburner is in a position of power to select items of key interest, e.g. Bernard's results, SO by Suzi)
- Judy will do GO for course at Woods Hole in November
- Midori will be doing 3 meetings
  - E-biosci
  - NetTab meeting – agents in bioinformatics (www.nettab.org)
  - Ontologies for Biology (European Science Foundation) – Heidelberg, Germany

- FANTOM, part of why David generated rules for a GO-slim
-
- KDD Cup 2002, this year will have both a FlyBase corpus and also and SGD corpus, the attendees are often from corporations and the results are rarely available to mere mortals, but may available to the AZ's of the world

Publications:

- Chris on GO database
- GO publication for Current Protocols – Judy and Midori
-

  Blake, J.A. and M. Harris (submitted) "The Gene Ontology (GO) Porject: Sturcutred vocabularies for molecular biology and their application to genome and expression analysis" in *Current Protocols in Bioinformatics*, Brazevanis, A, Davison, D., Page, R., Stein, L. and Storma, G., eds. Wiley & Sons, NY

- Matt's in Current Protocols in Parasite Genomics
- Evelyn: 1 to Genome Research (interpro2GO mapping); Bioinformatics article on InterPro; SwissProt article to Genome Research
- Midori – 2 requests
  - briefings on Bioinformatics
    - thought to be neither time nor cost effective to put such an article in such a small journal, so opinion against accepting either of these at this pint in time to avoid saturating market with the same thing again
  - Current Drug Discovery,
    - sort out gene nomenclature mess…
    - trade journal for portion of pharma industry

Website stuff

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

Amelia's website proposal

- good start on content reorganization

- suggestion to remove link to EP-GO browser
- statistics on hits:
  - 951 to AmiGO
  - 91 to MGI
  - 52 to EP-GO
- conclusion?

<u>Hinxton GO meeting</u>

- Genome Informatics (GI) meeting 4-8th September
- GO Users meeting September 9th
- GO Consortium meeting September 10-11

- users coming to GI, can extend housing an extra night, Users not coming to GI can refer to page of suggestions for housing, travel
- Consortium Members, possible to extend housing for one night via registration page, may be possible to extend housing via another mechanism, Consortium meeting will be in Cambridge rather than in Hinxton
- Suzi strongly encouraging attendance at GI, deadline for abs is middle of June
- Structure of User's meeting
  - Midori's thoughts, thinking of still having talks, but also having poster sessions, panel discussions
  - workshops with John, Chris software stuff
  - advertise such types of contents

Action Item 25 – Hinxton meeting (Michael Ashburner)

- **a:** find venue for 10-11 meeting
- **b:** get a Manchester person down to talk about DAML+OIL

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

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

Action Item 26c– Hinxton Users meeting (Midori) - mailing to go-friends list asking about desired content/attendance for User's meeting

<u>GO meeting after Hinxton</u>

- proposal to have it in John's home town in St. Croix, Virgin Islands
- hotels not too expensive

- TIGR – better in the spring, not winter

- late January - arrive on Friday 24th, meeting 25th-26th, leave on 27th January 2003
  - no Users meeting
  - pending quote from John

Action Item 27 – quotes for Virgin Islands meeting proposal (John Richter) - will get quotes and send to list within the next week

MGI Gene Ontology Progress Report of May, 2002
A GO-slim from a biological perspective (page 1/2)
by David Hill

Cellular Component
1.) **non-structural extracellular**: extracellular EXCLUDING extracellular matrix
2.) **extracellular matrix**: extracellular matrix
3.) **plasma membrane**: plasma membrane
4.) **other membranes**: (membrane EXCLUDING plasma membrane) OR (membrane fraction NOT plasma membrane)
5.) **cytosol**: cytosol OR (sarcoplasm EXCLUDING (sarcoplasmic reticulum OR junctional membrane complex))
6.) **cytoskeleton**: cytoskeleton OR microtubule organizing center OR spindle OR muscle fiber OR cilia OR flagellum (sensu Eukarya)
7.) **mitochondrion**: mitochondrion
8.) **ER/Golgi**: endoplasmic reticulum OR ER-Golgi intermediate compartment OR Golgi apparatus OR transport vesicle OR Golgi vesicle
9.) **translational apparatus**: eukaryotic 43S pre-initiation complex OR eukaryotic 48S initiation complex OR eukaryotic translation initiation factor 2B complex OR eukaryotic translation initiation factor 4F complex OR nascent polypeptide-associated complex OR signal sequence receptor complex OR ribosome
10.) **nucleus**: nucleus
11.) **other cytoplasmic organelle**: acidocalcisome OR cytoplasmic exosome OR endosome OR glyoxysome OR lysosome OR peroxisome OR vacuole
12.) **other cell component**: cellular component NOT (1-11)

Molecular Function
1.) **defense/immunity protein**: defense/immunity protein
2.) **cytoskeletal protein**: cytoskeletal regulator OR motor OR structural constituent of cytoskeleton OR structural constituent of eye lens OR structural constituent of muscle OR cytoskeletal binding protein
3.) **transcription regulator**: transcription regulator
4.) **cell adhesion molecule**: cell adhesion molecule
5.) **ligand binding or carrier**: ligand binding or carrier
6.) **ligand**: ligand
7.) **receptor**: receptor
8.) **other signal transduction molecule**: signal transducer EXCLUDING (ligand OR receptor)
9.) **enzyme**: enzyme
10.) **transporter**: transporter
11.) **enzyme regulator**: enzyme regulator
12.) **other molecular function**: NOT (1-11)

MGI Gene Ontology Progress Report of May, 2002
A GO-slim from a biological perspective (page 2/2)
by David Hill

Biological Process
1.) **cell adhesion**: cell adhesion
2.) **cell-cell signaling**: cell-cell signaling
3.) **cell cycle and proliferation**: cell cycle OR cell proliferation
4.) **death**: death
5.) **cell organization and biogenesis**: cell organization and biogenesis
6.) **protein metabolism**: protein metabolism
7.) **DNA metabolism**: DNA metabolism
8.) **RNA metabolism**: RNA metabolism OR transcription
9.) **other metabolic processes**: metabolism EXCLUDING (DNA metabolism OR RNA metabolism)
10.) **stress response**: stress response
11.) **transport**: transport
12.) **developmental processes**: developmental processes
13.) **signal transduction**:signal transduction
14.) **other biological processes**: NOT (1-12)