GO Consortium meeting Minutes: 23-24th September 2007

Sunday, September 23, 2007

Introductions (chronologically)

2007 Dmitry Sitnikov MGI, Seth Carbon BBOP

2006 Jim Hu E. coli, Susan Tweedie FlyBase, Trudi Torto-Alalibo PAMGO, Donghui Li TAIR

2005 Ben Hitz SGD

2004 Doug Howe ZFIN, Ruth Lovering UCL

2003 Jen Deegan GO, Emily Dimmer GOA, Alexander Diehl MGI, Mary Dolan MGI, Karen Eilbeck SO, Petra Fey dDB, Ranjana Kishore Wormbase, Pascale Gaudet dDB, Victoria Petri RGD, Kimberley Van Auken Wormbase

2002 John Day-Richter BBOP, Eurie Hong SGD, Tanya Berardini TAIR, Amelia Ireland GO

2001 Rama Balakrishnan SGD, Michelle Gwinn Giglio TIGR, Harold Drabkin MGI

2000 Rolf Apweiler EBI, Val Wood GeneDB, S pombe , Rex Chisholm dDB, Chris Mungall BBOP

1999 Midori Harris GO, Kara Dolinski PU, David Hill MGI

1998 Suzi Lewis BBOP, Michael Ashburner, Mike Cherry SGD, Judith Blake MGI

Progress Reports

Next year the GO Consortium effort will celebrate its 10th birthday.

2007 Progress Report for NHGRI due Jan. 1, 2008

These reports will review accomplishments to date.
We are using the itemized list of sub-aims from the grant to organize these

Aim 1: We will maintain comprehensive, logically rigorous and biologically accurate ontologies.

Ontology development

Ontology Development 1 - Midori Harris

All content meeting related changes documented on Ontology Development Wiki.

is_a complete was almost finished last meeting, but is now done and a system is in place to make sure it remains so.

Three high level terms need to be disjoint – cellular process, multicellular organism process and multi-organism process. General notes on is_a complete:
http://gocwiki.geneontology.org/index.php/Isa-complete_BP

Topics also overviewed; priority list on main Ontology Development wiki page.

Content meetings have been held for:

- Cardiovascular Physiology:

  http://gocwiki.geneontology.org/index.php/Cardiovascular_physiology/development

- Muscle Development:

  http://gocwiki.geneontology.org/index.php/Muscle_Development

Other topics:

- Transporter activities; extensive work via web conferences:

  http://gocwiki.geneontology.org/index.php/Transporters

- Medium-scale content changes:
  - synaptic plasticity
  - RNA processing:

  http://gocwiki.geneontology.org/index.php/RNA_processing

Michael A: Question IMG-to-GO and FIGS-to-GO mappings?

Jen D, Midori H: The IMG to GO mapping is mostly finished. These items are waiting for Jane to return.

Chris M: Mappings between the BP and MF terms still need to be done.

Judy B/Suzi L: wiki is a valuable resource, however it can get muddled sometimes – managers should keep track.

Alex D: If you add new large section you should send out a general email.

ACTION ITEM 1. Tutorial on wiki discipline (assigned to Jim Hu).

Rex C: In addition, there could be a group of wiki experts formed, who people could contact for advice.

Ontology Development 2 - David Hill

1) Taxon and sensu

"Sensu" confused users and curators, and editors became lazy in its implementation and accurate definitions were not created. Sensu terms have been renamed, merged or obsoleted (how many?) in collaboration with domain experts.

> **Note added after meeting**: We would have to run obodiff to get counts for renamed vs. merged vs. obsolete, but we started in April with 229 'sensu' terms, and there are now 80 remaining in the live file. Of these, several are sorted out in the 'fruiting_body.obo' file in go/scratch/, and the remainder (about 60) are listed on the last sensu meeting notes page:
> http://gocwiki.geneontology.org/index.php/Meeting_Notes_3

Definitions now need to state how a process occurs differently in the different organisms. If it is impossible to state this, then child terms will not be created. In future, term requests need to include reasons how a process occurs differently in different organisms.Synonyms containing the sensu information are kept for these terms.

The general consensus at the meeting seemed to be that rather than create long convoluted term names, we would still be allowed to create a term such as plant-type vasculature as long as the definition clearly differentiated the terms.

Function-Process Links

Chris M: These mappings are complex. Waiting for OBO-Edit 2.0 for help on cross-products.

2) Regulation

Regulation Main Page:
http://gocwiki.geneontology.org/index.php/Regulation_Main_Page

GO will soon add a new relationship, 'regulates'. Regulation-of-process terms will then be changed from part_of the process to regulates (for example, 'regulation of metabolism part_of metabolism' will become 'regulation of metabolism regulates metabolism').

During the is_a-complete work, three top-level regulation terms were added to represent three categories of biological regulation: regulation of molecular function, regulation of biological process, regulation of biological quality.

Chris has generated a report (go/scratch/regulation-of-non-process.txt) of all descendants of 'regulation of biological process' where there is no term for the process being regulated. David Hill is going through the report (not as bad a task as he'd feared), and has found that the violations fall into three categories, corresponding to the three parts of the Regulation Worksheet:

http://gocwiki.geneontology.org/index.php/Regulation_Worksheet

> Part 1: The regulation term describes regulation of a molecular function or a biological quality, so the term is o.k.
> Part 2: The regulation term is a legitimate subtype of its parent, but a more specific process term isn't required.
> Example:
> GO has 'regulation of transcription involved in forebrain patterning' and 'regulation of transcription', but not 'transcription involved in forebrain patterning'. 'Regulation of transcription involved in forebrain patterning' is:
>
> - Part_of forebrain patterning
> - Is_a regulation of transcription
> - 'Transcription of forebrain patterning' is not necessary -- it is essentially the same process as transcription

This term will inherit the regulates relationship through its is_a parent and will regulate 'transcription'. It will remain a part_of forebrain patterning since every instance of this process is a part_of an instance of forebrain patterning.

> Part 1: Actual problems of various kinds; David has made suggestions about how to handle them, which everyone should check -- especially the ones with question marks.

Chris M: There are problems with cross-products, and would be easier if the parent terms did exist.

David H: This will be resolved once the parent terms do exist.

David H: Concern about consistency in regulates relationships. In some cases, negative and positive regulation of a process are part_of the parent and in some cases they are is_a of the parent. We need to be consistent about this.

For now, negative and positive regulation of a biological quality are a special case. When you are regulating a biological quality, the regulation is a balance of the positive and negative processes. Therefore, the positive and negative children are part_of the regulation of the quality.

Midori H: Suggestion to use homeostasis terms for overall regulation of biological quality

Chris M: will look at relationships between cell types and GO terms: use as a guide to populate GO with missing terms.

Q. Val W: How existing annotations are affected by relationships change E.g. transcription intiation. may have annotated more granularly to regulation of transcription initiation when there is direct involvement. Topic for annotation discussion at some point?

A. David H, Midori H: The 'regulates' relationship shouldn't affect annotations. Basically the part_of relationships already exist and we will simply replace it with regulates. We are already annotating to regulates terms and it shouldn't change. What will be different is how we process annotations. We will be able to decide whether or not we should include regulates children.

ACTION ITEM 2 (ALL): Look at and comment on outstanding Regulation relationship items:

http://gocwiki.geneontology.org/index.php/Part_3

ACTION ITEM 3 (David H): Check whether there should be a relationship between pigment metabolic process and pigmentation

3) Information content analysis.

Collaboration with MIT/Harvard group.

MIT and Harvard got in contact with GO, were interested in measuring information content of a GO term. They looked at the number of annotations to a term related to its position in the ontology.

They developed a statistical algorithm to determine information content based on the assumption that if not many genes are annotated to a term it has a high information content and a term with lots of gene products annotated has a low information content

David, Midori and Jane then looked for outliers with respect to information content (finding terms that were either too specific, at a higher or lower level than they should be).Took higher level terms which had too few annotations compared to other things the same distance from the root and looked if they could be relocated.

e.g. pilus retraction was a direct child of 'cell physiological process' and was relocated to pilus organisation and biogenesis, so that it was at an appropriate level in the GO hierarchy.

Similarly lots of specific terms had a larger than expected number of annotation e.g. Olfactory receptor activity

Some of the annotation distributions between terms also just reflected biological differences e.g. cation and anion transport terms: there are more cation transporter genes than anion transporters, the two terms are at the same level in the ontology - as they should be. Therefore this analysis can only draw attention to particular parts of the ontology which a curator then can examine.

Q: John D-R: Is it possible to put this analysis into GOC tools?

A: Chris M: The analysis is already in database, can be used.

Alex D: This is something which can be repeated semi-regularly, but not to dwell on too much.

David H: This has been a very good collaboration experience, and had produced good contacts to continue relationships with. SR: We know of other groups that could also get in touch which are interested in this area as well - will get in touch with David Hill.

Judy B: annotations give power to these kinds of approaches. And until we have an annotation core we are restricting this kind of potential activity.

Ontology Development 3 - Chris Mungall

Wiki for ontology structure (should be merged with Ontology Development)

http://gocwiki.geneontology.org/index.php/Ontology_Structure

1. Mining Reactome links to link process to function – more after lunch.

2. Internal cross products can start to be created and maintained in the ontology. OBO-Edit 2.0 will make it easier to maintain these cross products.

New cross-product guide on wiki. Links to ongoing work on BP – CP cross products;

e.g. could link histone deacetylase complex to histone deacetylase activity (this type of linking is easier than creating BF to MF links)

http://gocwiki.geneontology.org/index.php/Cell_cross-products

Includes:

   Internal links (existing)

   External links (function to process links)

   External links (x products)

Links need to be treated with caution. Links are kept in a file separate to GO at the moment, as people could make erroneous propagation of annotations between the Gene Ontologies (i.e. just because someone annotates to a certain process, it does not mean they should necessarily annotate to the linked function).

3. contributes_to

Discussion:

Val W: People are using this qualifier incorrectly in annotations. Take histone deacetylase complex as an example, this is a very large complex with many molecular functions. Therefore one complex can be linked to many different functions.We should use contributes_to *only* in those instances where the annotator does not know which subunit provides a function.

Judy B: No, contributes_to can be used also when you *do* know the individual contributions of subunits.

Mary D: Often subunits which do not have a specific activity themselves are involved in enabling another subunit providing the activity.

Val W: But this does not hold for all complexes, we are using this qualifier in too many different ways.

David H: Often, if a subunit is knocked-out, the observer cannot tell if the subunit has a direct or indirect influence on the resulting phenotype. Therefore in addition often the 'contributes_to' qualifier is missing.

> *- discussion postponed.*

Internal cross-products

If cross-products were maintained in the GO directly, it would make life easier. Cross-products will be more manageable in OBO-Edit 2.0 where there are many features - can use a 'Cross-Product Matrix Editor' - can see the possible cross-product/GO combinations, parents and children of a term. This will help identify missing links in the DAG. In addition, there will be an ontology repair option, which can introduce these links, e.g. missing is_a links.

David H: We want to use this to go through the logic of the regulates relationship - as concern about ensuring consistency.

Chris M: will also look at relationships between cell types and GO terms: and can use as a guide to populate GO with such missing terms. ... more on cross-product logistics later.

SO Progress - Karen Eilbeck

Development:

March–>August worked with the J Thornton group - Gabi Reeves for BioSapiens project on protein terms, 96 new terms for polypeptides have now been added to SO.

With Mark Hathon (with Barry Smith) BioSmith, Buffalo – ongoing work on regulatory regions.

Content meeting in June, with HLA immunology community – looking for terms to describe variants. Added new terms, rearranging of SO, was very productive. Assigned work to Alex, who has nothing to report.

Collaboration with Arian at phyGo on mobile genetic elements for viruses. This is in parallel with work happening in GO.

Working on synonyms with Colin Batchelor, and over 400 new synonyms have been added to SO.

Release SO now every 2 months. Therefore there is a stable and leading-edge version for those interested.

Changing requirements for GFF3 - this not done yet.

Karen dropping down to 60% on this project.

COFFEE BREAK

Aim 2: We will comprehensively annotate reference genomes in as complete detail as possible.

Reference Genome Annotation Project - Rex Chisholm

http://gocwiki.geneontology.org/index.php/Reference_Genome_Annotation_Project

Aim 3: We will support annotation across all organisms.

see powerpoint file: GOC PU 2007final.ppt

Purpose: to provide comprehensive, robust collection of annotations for 12 genomes. These genomes have the most published data, have a genome database and experienced GO annotators. These high-quality annotations will be a resource for other groups to transfer to genes in their species.

Complete/comprehensive annotation includes measures of breadth and depth.

Breadth – every gene annotated.

Depth – gene annotated to the highest possible knowledge. If there are only a small amount of papers (5-10) then the curator should read all. If extensive then the curator should be selective, completion best assessed by a curator)

Target Gene Identification (Priority genes)

250 genes have now been targeted for curation. The target method has now been changed, targets are now (as of last month) selected based on disease type. Gene when mutated should contribute centrally to a disease phenotype(OMIM). This method has been generally successful, however there is now a challenge for mammalian groups with the increased literature load. Also a challenge for non-mammals - orthologs may not always be available (e.g. neurological genes in yeast). These challenges need to be balanced.

Ortholog Identification

Need to have a good set of orthologs.

Need to find ways of facilitating this work through tools, no obvious choice as yet.

e.g. InParanoid have problems in keeping pace and providing up-to-date sets. Would be good to have a ortholog set automatically provided which curators could then validate.

Software

Currently use Google spreadsheets for target lists and information on curation progress. However this is not robust enough and time consuming. A database will be developed to handle this data and requirements have been written up.

This will mean that the Ref Genome data is more structured. The database will provide a consistent use of identifiers, MOD association file loading, tracking when no ortholog found, and an automated response if a paper appears after a 'comprehensive date'. Sohel Merchant (left in July) wrote prototype. A new member of staff will start at the end of September to continue development.

Metrics

Annotation Progress

– see slide in GOC PU 2007final.ppt

Annotation Consistency.

Mary Dolan's tool for comparing annotations by looking at generic GOSlim

branches useful as different organisms are used in different experimental approaches and different levels of data are available in different organisms.

Eurie H: if there is an outlier in annotation consistency checks this might also indicate organism-specific data

Table View (slim showing each terms annotated for a gene) includes every term useful for curation and annotation consistency.

Ontology Development

Aim to have robust discussions on annotation and ontology development issues. Number of SourceForge requests from reference genome group in the hundreds over 16 months. There is an average of 10-12 SourceForge requests per month. GO editorial group doing a good job at keeping up with these. Existing requests are problematic. 411 terms. - It is important that curators label their SourceForge request as relating to a Reference Genome group.

Midori H: Can retrieve number of GO terms that have resulted from these requests by looking at the cross-references file: 411 terms from Reference Genome-marked requests.

Ruth Lovering's Metrics Document v3: see: HowToCaptureMetrics3.doc

Publicizing - Need to start publicizing Reference Genome work.

Annotation Outreach – Jen Deegan

Aim: To find new groups to join the GOC annotation effort, and keeping track of new groups annotating and writing documentation to help get groups started.

see wiki:

http://gocwiki.geneontology.org/index.php/Annotation_Outreach_group_reports

see: outreach_princeton.ppt

Jen described the scope and techniques of outreach effort. Showed an 'ontology ' of outreach effort. There has been much progress on grants.

Attending many regular conferences.

Less cold calling, it wasn't very successful. More luck tracking down the right person at conferences. Responding to invitations.
People going to meeting – report back information from willing people to Jen.

The SOPs have been tricky but are now on the public GOC website:

http://www.geneontology.org/GO.annotation.SOP.shtml

Michael A: this page is difficult to find. Action: this page needs to be reviewed and included in the next newsletter

ACTION ITEM 4. (Jen D): A reference to these pages should go in next newsletter.

ACTION ITEM 5 (Jen D): Add a link from outreach to the SOPs

There has been funding success - for the British Heart Foundation and AgBase grants.

Discussion:

Michael A: for new groups annotating, how many SourceForge requests are we getting? e.g. Aspergillus group should have requested new terms?
Suzi L: agree. As soon as an annotation effort really has started, the group often needs a number of new terms.

Jen D: For emerging genomes the main problem is finding funding to support an annotation effort.
Mike C: Need to determine if they are only doing IEA annotation, or whether they have the time to carry out manual curation.
Jim H: The process of making new term requests is not obvious
Val W: the SourceForge term tracker only goes to the GO list, so other groups not aware
Midori H: It is possible to add more e-mail addresses to this list.
Michael A: Not our job to source funding for new groups, it is the job of the individual groups.
Judy B: Supporting new groups is important, need to mentor groups and support them submitting new terms.

ACTION ITEM 6 (Jen D): investigate why terms requests aren't coming in, do we need to do things to make it easier, Regarding the SF tracker list and annotation list - who are on these lists and do other people need to be on those lists?

User Advocacy - Eurie Hong

Focusing on lines of communication, web presence, newsletter and mailing list.

Different users, new users, current users, power users

Most of the past year has focused on the lines of communication.

wiki User Advocacy main page:
http://gocwiki.geneontology.org/index.php/User_Advocacy

Rota of mailing list monitors.

Newsletters archived. Future news items page on wiki. Wiki or Newsletter ideas

Michael A: Wants ISSN for the newsletter.

ACTION ITEM 7 Eurie: Michael has sent URL for getting a ISSN for the GO Newsletter to Eurie, who needs to act on this.

Users meetings, we have a page of potential meetings on wiki. - used to target groups new to GO and help education. - have a workshop specific for microarray users (rather than an add-on to MGED)

Tools standards. (Needs to be cleaned up and categorised) - ideas for minimum standards for GO tools - send out list a month ago:

http://gocwiki.geneontology.org/index.php/Tools_standards

Production Systems - Ben Hitz

See: ProductionReport_GOC_PU_2007.ppt

Deployed 4 new linux machines 1 for loading, 2 for AmiGO production, 1 AmiGO development.

Production AmiGO now more fault resistant.

ACTION 8: e-mail Ben if you are not getting a gp2protein check for your database.

Go Database loading speeded up and now in testing.

Godb sequences – using gp2protein files. If possible do all sequences in your DB, not just annotated.

Assocdb fasta file – Header line massive – can be slimmed down?

Association file cleaning – All IEAs must have a with field.

AmiGO – Amelia Ireland

AmiGO enhancements and new search features demo

- Search result relevance implemented - most 'relevant' results are shown first

13

- Term and gene product search is now "intelligent" and AmiGO will automatically search all fields if it doesn't find a match.
- Term enrichment (also known as "GO Term Finder") and GO Slimmer (map2slim) functionality have been added to AmiGO. Both can use uploaded user files or data from the GO database.
- Downloads in OBO, RDF-XML and gene association format now possible

LUNCH BREAK

Action Items Review

This section moved to it's own page:

http://gocwiki.geneontology.org/index.php/Outstanding_Action_Items_from_17th_GOC_Meeting%2C_Cambridge_UK

Reactome - Peter D'Eustachio

See: Reactome_to_GO_GOC_PU_2007.ppt

Reactome can provide data to proteins that UniProt does not yet have manual annotations for most of this Reactome data is derived from experimental evidence identified from papers however unlike the GO annotation method, the types of experiments have not been recorded.

Discussion:

Emily D: GOA would love this data, but unless have a new parent 'Experimental' code, the best that exists is 'TAS'.

Suzi L: there is a use for a hierarchy of evidence codes. With an 'E' Experimental code as a parent of the IMP, IGI, IDA, IPI, IEP granular codes.

Peter E: Homolog sets used to transfer data between species is determined by individual experts, and transfer between orthologs AND homologs (where functionally similar)

Judy B and Suzi L: Reactome data is valuable. It is unacceptable to not be including it in GO and it is unacceptable that this data should have anything less than an experimental evidence code. TAS or NAS evidenced data are unacceptable also.

Peter E: current Reactome curation methods is to avoid unpublished data and Reactome curators also want to be opinionated about the published data, to the end that Reactome will reflect current expert opinion, and avoiding hypothetical theories. Only confirmed, accepted knowledge is included. There are 10 curators, only 2 of whom have previous experience in GO annotation, there is no budget to do GO annotation and no desire to teach curators about

GO evidence codes. Where there are multiple references cited in one Reactome page we don't always know which piece of literature applies to which annotation. 2000 genes have been annotated, with 4000 pieces of literature. It is not clear how many GO annotations this would convert to.

Suzi L: This brings up the question of what is the purpose of evidence codes? Why do we have the ones we have? Do users use them? (something to discuss tomorrow).

Pascale G: have evidence from users that they do care whether IDA or IMP codes are used.

Peter E: There is not always a 1 GO term to 1 publication relationship. Sometimes a GO term may have originated from the combined curation of many papers.

Eurie H and John DR: TAS annotations are valuable, and may be good to get the data in.

Suzi L, Judy B: This data is too good for TAS.

Emily D: Why not use a mix of codes depending on the GO term to publication ratio? For those instances where there is a 1:1 relationship of GO term to publication: use the new 'Experimental' code, for 1 GO term to many publications: use 'TAS' and cite the Reactome reaction web page as the source – this then acts as the reviewed document.

David H: Concerned about the proposition of a new 'Experimental' evidence code: might loose analytic power.

Judy B: Could Reactome curators go back and re-annotate those 4,000 papers and convert the codes to one of the GO experimental codes? This would only take 2 weeks to do.

Peter E: Not possible – Reactome have defined goals, we cannot afford to reannotate for GO. 75 genes/month is the absolute minimum annotations. We have our own grant objectives we must fulfill.

David H: GO curators could prioritize the reannotation of genes for which there is not much annotation available.

Rex C: Could the reference genome groups each take on a subset of annotations and re-annotate?

Emily D: Then the annotation would belong to the group that reannotated. We would be using Reactome data as a source, but the final annotations would be attributed to the group that provided the final annotations. Might not be the best use of resources.

Suzi L: Would accept 'EXP' for the 1:1 mapping of GO term to publication.
Q Val W: Any idea how many aren't covered by GO annotation already?

A. Peter E: No…
Judy B, Sue R, Emily D, Tanya B: the 'EXP' code would make life easier for users, for other integrations as well

ACTION 9: Reactome annotations should be available from GO by the next GO Consortium meeting. Chris, Alex, Jen and Ruth to be responsible. Add new evidence code EXP for 1:1 Reactome to literature, add all other Reactome with TAS to Reactome source.

 (continue discussions tomorrow concerning the point of evidence codes and the possibility of new parent 'EXP' code)

Protein Complexes: GO vs/ Reactome

Reactome complexes are seen as an entity, (i.e. a collection of proteins) whereas GO treats complexes as a subcellular location.

However there is also a blurring between the two for Reactome, especially when looking at large complexes.

Peter E: In our annotations, a cross-reference slot allows us to cite a GO identifier for the location (usually to the parent term of the complex). Reactome curators add the cc term that is most granular, and willing to generate SourceForge request for those missing

Judy B: talked to Lisa in Bar Harbor on complexes for Reactome. Concern about the active function tag to the active polypeptide.

Peter E: for a catalysis – any physical entity in a complex is given a GO term describing the activity, however the active unit, which mediates the reaction is labeled by Reactome. Can parse out which of the polypeptides had the catalysis functions and which are just associated – in most cases this is identified by experimental data. Although Reactome does not always search for the most granular Biological Process GO term, these haven't been applied consistently.

David H: there should be no problem mapping this data from Reactome, while the concepts in GO and Reactome are not equivalents this is not a problem as GO would annotate the same gene products as Reactome would.

Peter E: Ewan did have a concern about the 'contributes_to' qualifier – concerned that a significant number of end users would not always be aware of use contributes_to. But really this is the users problem. And they can strip out if necessary.

Jen D: users have suggested that GO could strip out annotations which use the contributes_to column (especially the NOT annotations) and these then could be provided as a separate file. As these can be dangerous to ignore.

ACTION ITEM 10: convert Reactome complex terms to GO terms (Chris M, Alex,D, Jen D, Ruth L)

Taxon and GO - Jen Deegan

see slides: axon and GO GOC PU 2007.ppt (using paper from Waclaw Kusnierczyk)

Originally Chris and Jen worked to loose sensu tags and redefining definitions and adding taxon links - However removal of taxon has been a problem. There are now 23,802 terms. Searching for terms is a time sink for users, - GO help has often received queries from users asking if there is a taxon-specific GO slim/subset of terms (e.g. plant-specific GO)

- In addition, Jen as outreach officer has found new MOD groups are unwilling to annotate to GO unless there is a slim available for them.

- GO language can be subtle. GO term names can now be complex now the sensu information has been removed. This would make GO terms easier to find and decipher.

- In addition, having taxon information in the GO helps error checking

- There are 3 types of relationships that could be applied to relate taxon to GO terms: 1. Is_relevant_to ` 2. is_only_in 2. applies_to_all

This taxon-specific information would be added into a separate file.

Discussion:

Judy B: Against including taxon information within the GO as we do not know all properties of a taxon. Taxonomic information is in flux also, we do not want a dependence on taxonomy in GO. We would be restricting ourselves if we did not make all terms available to all users. Could not instead users look at the terms that were used by a reference genome group to see what terms are appropriate for a particular taxon?

:: general disagreement from curators of this possibility.

Agreement that there are incorrect annotations which relate to taxon-specific properties:

Harold D: in the Fantom load – needed to remove incorrect mouse annotations

Val W, Harold D: InterPro2GO throw out problems. These could be identified by this method.

Val W: I perform monthly checks to ensure no inappropriate terms have come in at high level. This is time consuming, and this would help.

Pascale G: Would help sanity check annotation data

Val W: This species information doesn't need to be comprehensive to be useful for annotation checks

Eurie H: If this would help annotators, this information could be built into an annotation tool?

Ruth L: There are interesting concepts here, but does it need to be so complicated, would all taxons need to be included. Could we not instead just use just 10 high-level taxon identifiers.

Judy B: Instead, could not rulebase triggers be used. Efforts should be on annotation of literature rather than waste a considerable amount of time incorporating taxon information. We do not want to commit such a level resources to such a project especially as budgets are stretched presently. Again, concern about fluidity of taxon-specific information

Sue R: We should explore usage of GO slims.

Suzi L: There are risks in this kind of project, and concerned that this project would entail quite a bit of work and could also be misunderstanding by users. Can we have a low-key evaluation.

John DR: A large-scale activity of this – is a bad idea. You would propagate garbage by accepting all annotations. Could use as just a framework by only using 10 top taxon id. – this would already help find problems.

Jen D: Agreed.

Alex D: Isn't this just a user education problem? Users need to take the time to understand the GO hierarchy, that you can search synonyms, definitions etc. Feel that user queries are symptoms of users not trying hard enough to work with GO.

Mike C: Could not afford to make this a big project, there are other developments in GO which need to be addressed

Rex C: Had concern about making taxon-specific assertions that are flawed. If these types of sanity checks or limits were automatically applied, we would loose the potential value of not looking into these, however this data would probably tell us something fundamental about biology, and loose the ability to investigate these.

Judy B: Classifications of taxon are based on phenotypes and not molecular data and many things are being found and taxons are being redefined. Prefers' is_relevant_to' Like the idea of flags/triggers to facilitate work, but wouldn't automatically exclude, as this data is important.

Michael A: While some taxonomy is changing e.g. in protista, it is unlikely that viridplantea or mammalian will move around so much.

Ben H: what fraction of problems would be solved if there were cross-products to taxonomy were included?

Jen D: It would solve some, it would help with the development terms.

Ben H: What would the time line be for taxon cross-products?

Chris M: This is much further down the line.

Judy B: Our main issue here is how to facilitate annotations in our groups. However but we are hung up on a suggestion from outside the group.

Chris M: Slims are much harder to maintain than these relationships would be.

Michelle G-G: When the prokaryotic subset was created, she was v much against. Instead of users looking at 20,000 terms, they are now looking at 9,000 – there is not that much benefit. Don't think new users need this, need to facilitate better ways of finding terms within the tool. For curators it might be useful for error checking, but not new users.

John DR: Although there is a big concern that you'd loose annotations because of these relationships, this would not be the case as the incorrect annotations would instead be brought to your attention – and visible to better investigate/ or improve GO. the rules could be fixed.

Ruth L: How would this data be viewed ? In addition, if a user does not understand a term then it really is a problem with the terms definition – instead the definition needs to be improved, this would be far more valuable than adding in an additional cross-link.

Jen D: Will be willing to carry out a small pilot version of this task in her own time. Would add 10 is_only _in relationships and use these and the annotations to check for errors in the annotations and the ontology structure.

ACTION 11 Jennifer Deegan: do a pilot project with a minimal set of terms, as an experiment and bring back results for next GO meeting

ACTION 12 David Hill: Make difficult sensu terms organism specific (biologist intuitive) (i.e plant vacuole, fungal vacuole). However GO

definitions will still be designed to be formal, not depending on species to define the term.

Monday, September 24, 2007

Broad Agenda

- Plans for immediate future (SL)
    - regulation
    - cross products
- Database report
    - schema changes
    - production
- GA files (MC)
- OBO-Edit (JDR)
- Evidence codes (MA)

Overview of cross products - Chris Mungall

David H: Everyone should look at table 3 in the wiki:

http://gocwiki.geneontology.org/index.php/Regulation_cross-products People should comment on this so we can implement these cross-products. As soon as implemented then Chris will be able to run the reasoner.

Chris M: In future do we continue to run the reasoner periodically or should we put the adding of cross-products into the curation process?

David H: From the disjoint experience in the biological process ontology - we have problems if we don't get GO editors to put the information directly in, this is far better than going back and cleaning up these links. There are 6-8 ontology editors.

Chris M: Midori H will need to get this group together and teach them how to add in cross-product information, via Webex.

External Cross Products wiki:

http://gocwiki.geneontology.org/index.php/Ontology_Structure

Cross-product guide:

http://gocwiki.geneontology.org/index.php/Cross_Product_Guide

Judy B: A concern that using SO is exposing GO to risks, as SO changes

over time, is it a problem for GO to be reliant on SO?

Chris M: Not a problem, we can choose to use SO as it is, it does not force us to change anything as SO changes. In addition, adding in cross-product information into the ontology can be ignored if you don't care about the relationship with an external term, its fine but if you do add in cross-product information - the GO will be higher quality. Cross products results in scratch directory. They will be evaluated and moved from scratch to one of the main ontology directories

David H: How would these cross-product ontologies be edited?

Chris M: Editors would in future need to load three files - gene_ontology_edit.obo, xxx, biological_process_xp_cell.obo and use the cross-product interface

Midori H: It's good to always load these three files - as concerns about consequences of different levels of changes in each of the three files consensus: every editor should load all three files.

Chris M: Cannot merge the three files - as each at different levels of maturity, also the file would get every large. Will start with the regulates relationships and then the cell ontology cross-products

ACTION 13: David Hill to organize Webex meeting to ensure all editors understand what they need to do when inputting cross-product information.

Chris M: Wait until OBO-Edit 2.0?

John D-R: Start the regulation work now? Start with the cell ontology once OBO-Edit 2.0 is ready?

Judy B, Midori H: Yes, this would give editors a change to start and get used to adding in this extra information and identify any issues that came up.

Discussion: gene_ontology_edit.obo file vs. original file

John D-R. Add OBO version number to the filename. Then use original name as release version of file.

Michael A. Mike C. Change 'edit' in curators version to 'pre_release' to better describe its use. Original file updated nightly by Stanford.

Chris M. We need to take versioning a little more seriously – impossible to

replicate analyses. How do we cite what version of the GO we use?

Michelle G-G: Do we hide pre release file?

Chris M: No, culture of using the latest file.

Dan B: not straightforward to find when people took data

wiki: http://gocwiki.geneontology.org/index.php/Vershionning_Proposal

ACTION ITEM 14: OBO file renaming (Chris M and John D-R)

Judy B: add a link to Wiki:

http://gocwiki.geneontology.org/index.php/Versionning_Proposal

On the best practices page:

http://gocwiki.geneontology.org/index.php/Best_Practises

SO and Chromosomal Location – Chris M and Karen E discuss offline.

Term Lifecycle - John Day-Richter

Term requests to instantiation in the ontology has been reported to be a bottleneck. Users request terms, and then need to wait for implementation to use them.

Proposed solution. Give users a temporary ID to work with when they need a new term. Create mini ontology file they can update all their annotations with the new term id.

However many terms are rejected as requests are inappropriate, so need to feedback to them outcome of term request? How? Everyone has some way of dealing with term obsoletion – therefore we can use obsoletion mechanisms to feedback to user. When request closed, use 'consider' or 'replace_by' tags to get correct term. The term is obsoleted in the users private ontology.

Discussion:

Pascale G, Rama B: New groups might not be able to handle this. How many new groups are requesting terms and need them straight-away?

Alex D: if new curators request a large number of inappropriate terms and then annotate to the temporary ids, this would end up irritating the curators. A peripheral group may not be able to handle this well.

Mike C: automation not a human friendly approach, this is not user support per se.

Midori H: How much more burden to track down terms to suggest and consider a replace_by?

John D-R: Most term requests a lot of work, might be easier to phone the person and do the request on the spot.

Suzi L: But this is an extra to personal communication. Shouldn't stop a group from requesting new terms where there is a need for a fast turn-around, do not want to discourage new groups.

Michael A: How does this fit in the with SourceForge tracker, should we release this? Or keep as an option for groups submitting large numbers of requests that we trust.

John D-R: The obsolete temporary terms would die after 3 months in your mini ontology, therefore if the user did not carry out the required updates they would lose their temporary id as well - this creates a large incentive to properly implement obsoletion handling.

John D-R: Seth has already done this – but sounds like we shouldn't release this publicly.

Judy B/John D-R: Not much support for this – put on backburner until we find a good project for this to be used on. Possibly to use for Reactome requests, but Midori pointed out that ALL Reactome requests (ONLY 17), had been dealt with very promptly (12 within 2 days only one took 3 weeks to resolve).

ORB (Ontology Request Broker) - Seth Carbon

Seth has implemented a SourceForge request tracker wrapper. Will be available in AmiGO where users can use it to directly make a SourceForge request. Both HTML and programmatic interface for SourceForge, available for any tool.

Seth: DEMO of this wrapper in AmiGO

- Users unable to find a term are given the option to go to a page which allows users to provide details and request a new term.
- A form is provided to add term name, definition, additional details, and (optionally) a SF ID. This is submitted to the SF tracker, gives a success ID, term added into tracker. Users can choose to put in a user name or if not the request would go into a general id bucket. The user

23

can then retrieve their terms with orb_default ids in OBO format.

Could not track submitters by IP address, as problems with firewalls and dhcp.

Could make a batch request - and an interface could be written so that an individual SourceForge ID would be returned for each request.

David H: Want a way of identifying users, temporary ids are worrying.

Emily D: Within the AmiGO form, it should be essential for a submitter to add their e-mail address, would ensure they are involved in the term-request process and reduce number of lazy submits by users/spam.

Michael A: Attribution section of the AmiGO form should be changed to 'E-mail Address', this should be required.

ACTION 15: Midori to work on the GO stanza specification required.

Michelle G-G: provide link to new term best practice documentation

John D-R: Use one batch tracker id?

Seth C: Perhaps generate SF ids using another system?

ACTION ITEM 16 (David Midori Seth) Deploy the part that created SF items based on a friendly webform, and would like to see an obo format in the SF item.

ACTION ITEM 17 Link to documentation on how to make a perfect GO term

Schema changes - Chris Mungall

SWUG:Database changes 2007

- Support for multi species annotation files
- (PAMGO to task releases files next month. Trudy to send examples to Chris to test this facility)
- Support for new properties column. Test data from MGI received (they use structured notes field, others should also send examples)
- Support in schema for taxon based queries, species, kingdom etc.
- GOOSE new interface to MySQL DB. Aimed at intermediate to advanced users. EBI mirror>5000 hits so far.

GOOSE

SQL query interface for intermediate to advanced user.

http://www.berkeleybop.org/goose

Provides example SQL queries example: Stale ISS assignments

There is a wiki page of example SQL queries - can use to experiment and alter to needs - results in html or tab-delimited. GOOSE does not use the production database in Stanford, and a kill-limit is in place for problematic queries. Full version of the database is queried, including IEA data. Mirrors are updated daily, but annotation source updated once a month.

Q. Susan T: web services?

A. Chris M: Already a GO Perl API but will be providing web services and sparkle already ready

New architecture road map on AmiGO. More interactive components on front end.

wiki:

http://gocwiki.geneontology.org/index.php/SWUG:AmiGO_Architecture_Road map

Seth and Amelia have been refactoring the server based code. Roadmap to transition AmiGO from Perl to Java. Re-use existing OBO-Edit code, mature and robust. Therefore saving development time in future.

Renovated GO database info page.

ACTION ITEM 18: Amelia link GOOSE from front page - DONE

Gene Association Files - Mike Cherry

SGD wants to have 2 files – one manual, one IEA.

SGD would like to provide IEA annotation predictions. However concerns over ignorant users analyzing GO data and using existing IEA annotations to create circular annotations. What should SGD call this file?

Chris M: need consistency if SGD do it - then we all should do it.

David B: want to make it clear to researchers that they should use correct GO annotations.

Michael A: propose that groups submit files as normal, but that a filter is

installed for all GA files which will partition them into IEA and non-IEA files. On the annotations download website and ftp sites there will be too files.

Emily D: concern over proliferation of files. DB: Need something that will tailor files to user requirements (taxon, NOT, non-IEA), advanced interface to do this on-the-fly

Jen D: Help education of IEA. Important that GOC members who review papers are aware of this problem - place this information onto wiki.

Jim Hu: This reviewing information should be made public, should be a page both for those writing and those reviewing papers. SR: information on evidence codes is hidden in the GOC site, should highlight location of this information on the GOC front page.

ACTION ITEM 19: Mike Cherry to write Gene Association file filter script

ACTION ITEM 20: Chris Mungall: Create more advanced interface to download custom files by versioning

OBO Edit Working Group - John Day-Richter

- About to release 2.000 beta-14
- 89 bugs fixed.
- OBO Edit toolkit now used in Phenote.
- Reasoner much faster. Edit in real-time with reasoner on.

DEMO of OBO EDIT new features

- Auto-complete
- Advanced searching for power users, Boolean querying
- Advanced sub query feature
- Docking panels to personalize interface.
- Graph based editing updated automatically
- Wrench icon for every panel to set up personal preferences, filtering, view options etc.
- Create new terms and relationships in graph editor by drawing
- Graph overview preview
- Graph DAG Viewer

- Spell checking
- External contributions from..... CJM

Image:OBOEditWorkingGroup GOC PU 2007.ppt

Image:Term Requests GOC PU 2007.ppt

Discussion about availability of predictions - David Botstein

As more information about genes becomes available, the hope has been that this would in turn evolve high-throughput methods to find out more about genes and interactions and pathways and that this data would be determined and incorporated into GO. While this has happened to some extent, it has not been at the rate hoped for. There is no iterative process.

The ideal would be a list where curators would be promoted to check that validity of an annotation prediction. In reality, because of the vagaries of IEA data it has been difficult to validate data.

When people have looked for statistical links between genes to looks for possible associations, most predictions turned out to be good. They found evidence for this that wasn't currently included in GO (i.e. un-annotated but information present in literature)

Would like the GOC to start using the suggestions coming from the community.

Judy B, this is a priority issue. MGI has experience with Ken Pagan's group regarding mouse genome. Does take time to have an interactive relationship. There are genes which only have IEA annotations, this must be a priority set for annotation, they are key for curation and need to be prioritized.

David H: did this exercise with Fritz Roth's dataset fell into 3 categories: correct annotation should clearly be made enough circumstantial stuff to make this annotation, but not tested from outer space

David H, Emily D: this exercise takes a lot of curation work

Mike C: FRIZ, OLGA RC: have a grant to look at BioMediator - using expert GO annotation to validate predictions DB: could be something to build into tools.

Conclusion "Need more curators"

Sue R: Users group to focus on predictions?

David Botstein: Some are one-off; others are systems which should be semi automation. Does anything arise from the algorithm which isn't obvious from reading the paper? Use the best of the methods routinely

Judy B: Should get reports from people who have done these types of collaborations

ACTION ITEM 21: Predictive Activities. Collaborations with external groups. Reports into next GOC meeting as to these kinds of activities. (assigned to: ALL)

Suzi L: something we build in to the long term. If GO becomes responsible for running SW/ limiting.

David H suggestion, run on reference genome gene of the month

Need to leverage the groups who are doing these things

Jim H: Suggested making a repository for predictions. Set up a place where people can dump their results, and we will look at them.

TOUR OF LEWIS SIGLER INSTITUTE

Group Photo

LUNCH

Annotation Evidence Codes - Mike Cherry

Need to finalize the proposed evidence code documentation. There have been a huge number of e-mails on this subject.

Rama B: There is a draft web page ready, with the majority of text and examples agreed upon. However there are some sections in read, where input from the GO Consortium is needed.

Chris M: Assume that what is not in red is okay, and lets move on and get it onto the web site.

Pascale G: Evidence codes documentation is too long and complicated. Would it be possible to also come up with something simpler for users - and leave the long documentation for curators?

Chris M: evidence code documentation should be for curators, but users do read it as well.

Eurie H: We should have an abbreviated version and move the detail onto the GOC wiki, as this level of detail swamps users.

Suzi L: Seconds Eurie's motion

ACTION ITEM 22: Evidence Code group to make changes to Evidence Code documentation so that details for curators and summary for users are separated.

Decision tree pdf for evidence codes has appeared. No one has seen before - from Karen? Looks good. GOC needs to look through this.

Discussion: Revisiting the question "What is the purpose of evidence codes?"

How are evidence codes used by curators, biological users, informaticians

Users get an idea where the GO annotation came from

Val W: Curators can use to evaluate conflicting evidence from other species to make the best ISS inferences based on the available data. Might not want to make an ISS annotation based on IMP evidence.

Sue R: for functional inference, manual curation used as a gold standard for bench marking. For instance, if they are making inferences based on expression, they should remove IEPs. When inferring from homology need to exclude annotations made from homology. It is important to stop ISS and IEA annotations becoming circular. More confidence is achieved with varied evidences rather than one type of assay.

Eurie H: Users are using evidence codes as a confidence level rating.

David B. Future of computational analysis depends more and more on the evidence codes and that some evidence is stronger than other.

David H: To quote documentation that there is no quality measure in evidence codes is patently false - all users use evidence codes in this respect.

Eurie H: Old documentation said evidence codes can be used to "Evaluate the reliability of an annotations" this is the original intent of the evidence codes. There is a natural hierarchy.

Judy B. Experimental codes have been working well. Debate mainly outside

of the experimental evidence codes.

Rama B: would be use for to query communities on their awareness of evidence codes – do they know what they are, what do they can use them for? Also would help regarding the proposal of expanding evidence codes – we could then get a feel for what would be of benefit ?

ACTION ITEM 23: community outreach regarding evidence codes (outreach group)

Michelle G-G: Many organism don't have the literature to draw on, all meaningful annotation is sequence-based methods. Many (99%) of genes have no literature.

Rex C: RG goal use data based on experiments. Importance of evidence codes is paramount. Philosophical reason, provides a broad a base as we can with the groups that have experimental data for the groups that don't

Michelle G-G: The vast majority of organisms do not have literature resources to draw on. Therefore if we want annotation we need to draw on other methods. There is a vast prokaryote population with no annotation.

There are important distinctions to be made in ISS. There is a huge spectrum of quality in ISS, i.e. if someone only takes the top Blast hit - that's a low quality ISS. There are orthology based methods ISS, SnoRNA predictors, signalP, TMHMM, tRNA scan. Purely sequence analysis should be ISS

Mike C: for experimental annotations there is also the problem of a large different in the quality of published experimental investigations

David B: When we talk about inference, you need to ask what is this data being inferred from? The source of that data is all important, need to make a trail from ISS statements to the source of that idea.

Suzi L: Important to ensure that in the 'with' column we have information on what the ISS data originated from, as need traceability.

Judy B: Declaration of orthology for mammalian groups provides a basis. Concerned about extension of ISS

Rama B: RCA code was initiated by SGD for annotation from papers which use a combination of methods.

Rex C: For non-mammalian organisms, where there is not much experimental data, it's important that a curator can see similarity without having to define it

as an ortholog. Curators should be able to make a reasonable inference.

Ben H: Question comes down to which term; don't need strict orthology to infer protein kinase.

Judy B: Yes, but the overriding theme is that we used one ontology applied by many different groups. These annotating groups must now agree standards - we don't want generic sequence similarity. Ortholog is really important. When there is credible evidence that there is an ortholog this should have a separate ISS code. And if a gene has a functional annotation to transfer, as there is a well known active site then this is ISS.

Michael A: If ortholog tables could be trusted, ortholog evidence code can be computed

Emily D: In GOA ortholog sets provide basis for an IEA prediction method. Curators can then manually infer functional equivalence by looking closer and assigning manual ISS annotations both for orthologs, paralogs etc.

Rex C: Orthologs are a more complex characterization of a sequence alignment should be able to put a sequence in the with column. Sometimes ISS unclear, which is the ortholog? If you can put something in the with column. Use ISS otherwise RCA. If the method is computational, requires building of model, whole bunch of approaches, computational analysis.

Sue R: TAIR uses granular children of main evidence codes.

Mike C: Against proliferation of evidence codes. Could have orthologs as IEA.

Judy B: Orthologs not just based on domain structure, but whole protein.

Harold D: There appears to be two flavors of ISS, but the distinction is what goes in the 'with' field. How much future away is the result of a tRNAscan than a measure of sequence similarity? It is misleading to have different flavors of the ISS code both for users and curators.

Mike C: We rejected many evidence codes, so there has to be 'collections'

Emily D: Isn't the finer grain detail provided by the GO reference collection?

Jim Hu, agreed with Rex, ISS based on orthology, overall partial/paralogs/families TMHMM fundamentally not evolutionary arguments

Val W: tRNAscan could be ISS or RNA, The tRNA and snoRNA predictors use multiple methods and could be RCA - as computational analysis by a

combination of methods.

Ruth L: Ortholog data will change from time to time, IEA data will change.

Rex C: Does not wanted to be restricted in ISS on the lack of evolutionary relationships. If you cannot put a sequence in the 'with' and the analysis is computational that requires a model involving a bunch of processes then we cannot put something in the 'with' as a sequence. This could be a good operational approach.

Sue R: Not adding more evidence codes is unrealistic. And adding evidence codes one at a time is primitive and its time people think about the whole picture.

Suzi L, John D-R: Agreed. Evidence code ontology should be looked at. This does not mean increased complexity for users; users can choose which evidence level they look at data.

ISA overall sequence alignment (OSS) [seq:ID]

ISO - orthology data [seq:ID]

ISM - model, hmm or SCFG [with is optional, only used when identifiers to model]

David B: but slimming will not help naive users.

Ruth L: Moving ahead, a working group needs to make a decision and sort out something which works.

EXP as a higher node

People can do this without changing the way they annotate Would allow people to download data with the relevant evidence codes

Settled on the following proposal:

ISS

    ISA requires sequence ID in with field
    ISO required sequence id in with field
    ISM

EXP (new grouping term for experimental evidence codes)

    IMP
    IGI
    IPI
    IDA
    IPI

RCA a more complicated method

Proposal ECWG to make new evidence code hierarchy.

ACTION ITEM 24: Evidence working group to make new evidence code hierarchy in the context of what has been discussed this afternoon.

Implement richer number of evidence codes.

Query communities about evidence codes. What would benefit them?

Michael A bequeathed the evidence code ontology to Sue Rhee

There was a short discussion about whether IGI should be used when only one gene was mutated. Consensus was that it should not

ACTION ITEM 25 (Evidence code committee): Revise evidence code documentation so that a mutation in only one gene can only be IMP (protein localization IGI example)

ACTION ITEM 26 (Curators): Check whether you have used IGI in this way and update annotations

Short discussion on usage of the 'with' column for NAS. All agreed that the option of adding a GO identifier in the 'with' column when using this code should remain (as previously agreed)

Short discussion on whether only ND is allowed for root nodes. Was agreed that ND should only be allowed when annotating to the root nodes. Documentation needs to be clarified. This represents a status item.

ACTION ITEM 27 (Sue R, Michelle G-G, Rama B): Put evidence code proposal in the context of what we discussed today

ACTION ITEM 28: Update evidence code decision tree in response to today's discussion on evidence code usage (Jen and EV Code WG)