GO Consortium meeting minutes

Sunday afternoon

Minutes: Yasmin and Emily

Judy Blake
The new GO grant will start 1st March. This meeting will focus on the activities and the GO mission in the next grant period.

We will be talking about the GO-NIH funded activities, but also those that many other funding efforts contribute. It is also important for us to understand how the GO annotation stream fits into the stream of other annotation projects in the different contributing groups.

Current GO annotation status

Suzi Lewis
Presentation of metrics of GO annotation over the 5 year perspective of contributions by different groups, data supplied by Amelia Ireland.
Variability between and within a group annotation contributions is high over time and affected by changes within groups - for instance due to a new genome assembly, e.g. WormBase, changes to a group, e.g. PomBase.
EXP-evidenced annotation gradually increases over time, however the annotation rate has been constant.

General question - how can curators can be helped to become more efficient?

The GO PIs will be analysing this data in further detail and will eventually make some of this publicly available.

Paul Thomas and Mike Cherry
Summary of responses to the GO survey from annotation groups
Full responses from each group are available from the GO wiki's meeting agenda page.

Question 2

2A-C. How many different curators are working on GO annotations? How many FTEs (full time equivalents) does this represent? (If you have 2 curators working 50% time each on GO, then it would be 1 FTE total.)
Are there any curators in your group who contribute to GO ontology development?
Does your GO annotation effort benefit from any dedicated software support (from any funding source)? How many FTE?

21 groups replied,

Annotation effort: 19 FTEs + UniProt 8 FTEs.
Ontology development embedded in annotation groups: total of 2 FTEs
Software development: 6 FTEs

Total: 36 FTEs working on the GO project in annotation groups. Therefore we have a large resource to draw upon for advice on best practice.

2D. Is GO curation integrated with other curation processes (does a curator specialize in GO curation tasks, or do they do capture GO as a component of overall curation task).

19 replied - 16 carried out integrated annotations within their groups.

2E. How do you prioritize GO curation relative to other curation tasks?

6 groups prioritized annotation based on papers, 6 based on proteins. 4 had no prioritization method

2F. Please summarize all of the curation and data representation tasks going on in your group (please include taxa coverage, domain coverage, source of data)

most(14 groups) annotated between1-4 species, 7 groups >5 species.

Judy: while we're trying to facilitate the production and tracking of annotations, we (GO PIs) need to know if/where this is clashing with other activities.
There will be a need for improved reporting of changes in the next grant period - metrics, showing progress, QCs

Rolf: concerns of counting effort wrt grant reporting, PIs should only count people funded on the grant, not consider all annotation efforts.

Mike: we also want to gain an overview of the project, not just those GO NIH funded activities.

Suzi: annotation stats will be counted automatically and webpages will be created to display and track changes in annotation files over time.

Paul S: such stats are useful for internal understanding of priorities and qualities of annotations.

Emily: would be nice to include along side these numbers, an often-updated summary/blog regarding the changes in the annotation effort over time, e.g. the focused annotation efforts being carried out by the group, new QC checks being integrated. So that users get a better understanding of how the GOC annotations changes over time with regard to contents

Suzi: measuring quality is difficult, but important. Need to fold in reviews for pathways etc.

Pascale: if this data public, we need to be sure as to what data we want to present to users.

Question 3
3. Here we seek to understand your process for GO annotation
5 replied, 4 stated that the pipeline is manual. It would be good to learn from each other regarding how we each carry out annotation.

3A. What software tools do you use for GO annotation? Include any curation interface tools as well as paper triaging, text mining and sequence analysis tools.
20 replied - 9 use in-house software. Therefore much to share in the Consortium for both software and experience.
7/21 used Textpresso. This is therefore a highly popular tool.

3B. How do you identify and prioritize which papers, genes, etc. are targeted for GO annotation?

Some groups focused on annotating gene families/target lists, many prioritized based on date of the previous curation effort.

3C. Do you regularly make both literature and inferred (e.g. ISS, IEA) annotations?

Most curators individually create their ISS set.

With regards to IEA annotations:
- 4 groups don't use any
- 6 groups use UniProt IEA annotation sets
- 3 groups carry out IEA pipelines in house.

Therefore already a lot of standardization on how this data is being produced

4A. What gene function-related information do you currently, or do you want to, capture using a controlled vocabulary that you currently CANNOT capture with GO terms?

Phenotypes most frequently annotated, then interactions after which allele annotation. Only 2 groups described curation of regulation mechanism. Few groups mentioned curating pathways. How can we learn from each other?

Judy: in some groups, e.g. MGI - there are other divisions of specialised curators carrying out such annotations, therefore this overview might not be complete
Mike: this summary is solely based on the information provided in the questionnaire responses.

Paul T. (see slides presented)

3F What is your process for creating a GAF file for submission to the GOC?

everyone has in-house scripts for creating and depositing files, some groups have QC checks. A few groups submit via UniProt

Some groups integrate annotations from external groups for same species - authoritative

species owner groups.

3G What types of references/sources do you cite for your annotations? Are there other types of information source that you would like to use for GO annotation?

- - quite a range of id types used to support annotation (pmid, doi, books, urls, Agricola, older refs not in pmid)

3H. Do you currently provide annotations that include the ANNOTATION_EXTENSION or GENE_PRODUCT_FORM_ID information? If not, have you any plans to start releasing such information in the next 12 months?

7 out of 19 groups : use the annotation extension field
5 out of 19 groups: use the gene product form id field.

3I. Has your group decided not to produce or release any of the supported types of GO annotations? (e.g. IEA or IEP-evidenced annotations, protein binding, column 16) and if so, why?

Main types of annotation rejected for curation:
-IEP-evidenced annotation, protein binding descriptions, NAS/TAS annotations, use of the colocalizes_with qualifier.

3J. What are the obstacles to (or rate-limiting steps in) your GO curation and submission pipeline, aside from curator resources?

- selecting appropriate GO term.
- the time taken for finding terms.
- waiting for new terms
- developer time in response to changing req/stds
- literature triage, text mining
- exporting to GAF
- curator training
- Annotating HTP expts

Kimberly: new term requests should be dealt with faster and this will be incorporated in new annotation too

Varsha/David: Raised a point regarding the delay in waiting for terms requested by SF. Waiting for this lag. Would like to access new terms faster via CVS in future.

4. In what ways can the GO Consortium assist groups?

How can we make the process easier - and reduce duplication?

* software support
    Curation software support
    Community annotation tool
    Guidelines/support in propagation of annotations
    Annotation QA support
* assistance in updating annotations after an annotation change.
    better communication at meetings and improved annotation guidelines.
    re-annotation assistance
* ontology development issues

Paul Thomas: should be better feedback on how to annotate if a proposed term is rejected

Paul S/David: scope of annotation changes. Some times specific terms obsoleted, other times there is a re-development of a node of GO, which causes obsoletions in many terms - e.g. transcription.
    Topic to be discussed further later in the meeting.
    large scale obsoletions should be discussed at meetings, small scale term obsoletions can be done individually
    Karen: for transcription, we were aware that large no of annotations were affected and documented in advance of obsoletions
    David: for smaller obsoletions: replaced_by tag. Could be done centrally.

    Jane L.: much regarding prokaryotic biology is missing.
In addition, plant secondary metabolic processes under-represented (Lukas M.), fungi also under-represented (Diane I.)

* Better comments and definitions to resolve ambiguity needed

* Curation software support requested:
    Curation interface – tools to help find GO terms – “curators who use this term also used…”
    Annotation of related genes in other organism
    Annotation of terms to
    Suggest GO terms from free text (text mining tools)
    Literature prioritization tools\extensions
    Community annotation tool

-Emily: interPro appreciate the feedback that they get from SF –but more could be given. They would also like to be notified when we have a focused annotation effort so that they can re-evaluate their mappings

4C. Are there other ways the GO Consortium could help your GO annotation and/or submission process?

Documentation e.g. GO-approved term usage
ISS checks and 'with' validity
centralized QC functions
increased feedback/collaboration for InterPro - further feedback and notifications on annotation efforts.
consistency across related genes

* community/users
* annotation expressivity and additional annotation information
expressing time- or condition-dependence, the annotation_extension field is useful for capturing additional specificity, Need to express time or condition dependence

David: this should go hand-in-hand with ontology development blockages
links to terms from other ontologies
annotation statements and chain of evidence from multiple publications

Judy: it s a challenge to decide our priorities - depth vs breadth issues. What are the bounds of the GO project? It is important to work out how to make this project more useful to our users. Resource are limited, so need to maximize our efficiency/usability.

David: we want to provide information fully: but do our users?
Pascale: different groups might want to have different directions regarding annotation expressivity according to the research being done on their organism.

Harold – as an aid of this – all people on GO help –can see what questions come in and could help to prioritize what to annotate.

Suzi– want to make a small working group to develop the annotation stats reporting pages: need about 5 people. Need to set up a framework and then do content. Annotation numbers would be good too to have on these pages

Paul Thomas: GP2Protein files.

This was originally a mapping file to map MOD gene ids to ids either UniProt or NCBI ids.

UniProt reference proteome files are available. These identifiers can be directly annotated if a group does not want to maintain a gp2protein.

Format of the gp2protein: All protein coding genes are represented by one canonical protein sequence per gene
Purpose
1) To allow GO users to find genes/gene products with annotations, including intelligent messages when no annotation exists for a known gene product ; used by AmiGO/GO database
2) To allow GO phylogenetic annotation to be applied as broadly as possible – used to make gene trees/ortholog sets ftp.pantherdb.org/ortholog
3) retrieve gene product sequence information.
4) allows users to map between MOD id types, if interested in a multi-taxon annotation

set.

History: Paul Thomas's group worked with several mods in 2007/8 to help generate compliant files. UniProt began to create files for this effort in 2009

Judy - MGI has issues with supplying a complete proteome for MGI – MGI uses VEGA mappings. GI's output to gp2protein is automated and not all ids have a UP id

Claire – we have all the VEGA sequences for mouse and hence equivalent UniProt ids. XP and NPs are covered too. Rolf volunteered Dan to do this who left after 9years. Eleanor is now working with Maria and Mark from SIB – this data is imported by Eleanor – via Vega ids.

Judy - MGI will discuss this further with Claire. Other groups should similarly contact UniProtKB if mappings are a concern.

Claire: UniProt help (http://www.uniprot.org/contact) is a good way of communicating with the group.

UniProt Reference proteomes - led by Maria Martin at UniProt, contributed to by Dan Barrell and Eleanor Stanley

- One release per year
- Covers 66 species
- All species represented by survey respondants
-

Chris - Is the gp2protein file fit for purpose? Should we be looking to use the GPI file?

Requirements for GO
All GO annotations must try to describe an identifier in the Ref proteome set, or provide as complete gp2protein as possible.

Non-coding ncRNAs
If your group annotates to nc RNA genes
Please create a supplemental mapping file - gp2rna – to mapp MOD identifiers to RNA entries in NCBI
This file should also contain all know ncRNAs in your genome.

Rolf – GOC should pay attention to future developments in the RNA community, there are plans to have a UniProt-equivalent for RNA.

Susan: neither the gp2protein nor the gp2rna files contain those which have no sequence attached (unlocalized genes. In addition, we need to have authoritative, stable written document for format changes.

Susan- need to update wiki and there needs to be a stable path to the latest documentation.
Emily – documentation should on a centralized wiki – groups need to be pointed to it. Currently documentation is incomplete and distributed in the wiki. It is a huge task to centralize it all.

Conclusion:

Paul: over the last 10 years GO has grown up. Big picture: Expecting GO annotation on a large scale – need to make sure it's maintainable etc….and we want to be delivering to our

users the best quality, relevant data.