Overhaul of Representation of Transcription in GO

Biological subject expert: Karen Christie Ontology development expert: David Hill

Ontology Development Phase

This project was converted to a high priority project due to synergy between several things:

1. Accumulation of a large backlog of old SourceForge items relating to major flaws in representation of transcription, especially in the Molecular Function branch.
2. Karen was invited to speak at EMBO Gene Transcription in Yeast (at their expense)
3. David was willing to work with Karen to provide expertise on appropriate relationships and ontology structure

In early May 2010, Karen and David began evaluation of the existing problems in the representation of transcription and transcription factor terms in all three ontologies and developed a number of outstanding questions. Karen and David held regular weekly phone conference calls to supplement email communication throughout the process.

In late June 2010, Karen attended the Gene Transcription in Yeast meeting with a poster on the transcription overhaul project and got lots of input from experts in the field. Based on the prior research and questions addressed at the transcription meeting, Karen and David worked up a proposal for changes, mostly in the Molecular Function ontology with a few changes in the Biological Process ontology, which was sent out on July 27th.

Comments from the GO Consortium on the initial proposal was generally favorable but indicated two major issues to rethink. One of these related to ambiguous usage of the word "promoter" in the literature. The second related to the decision made at the Geneva Annotation camp (June 16-18) that binding relationships would be implemented with the has_part relationship when binding was an integral part of a more complex molecular function. The requirement to use has_part relationships made it necessary to significantly rethink representation of the transcription factor activity terms in Molecular Function. The revised proposal sent out on August 24th.

Throughout the project, Karen did extensive reading of transcription literature including yeast, mammalian, plant, and bacterial transcription systems. David supplemented this with specific reading of mammalian literature. Jim Hu and Debbie Siegele were consulted several times and provided invaluable knowledge of the bacterial perspective. Julie Park and Harold Drabkin provided valuable input on the subject of reverse transcription versus RNA-dependent transcription. We also talked with Karen Eilbeck with respect to coordinating the representation of DNA sequence elements relevant to transcription between GO and SO. As annotators, both Karen and David were able to evaluate the literature both from the perspective of annotators as well as that of ontology developers.

To resolve a couple issues, we consulted experts in the field via email. Both times, the majority of the people we contacted responded and provided insightful feedback. Though somewhat time consuming, this was invariably useful when we felt unable to resolve different perspectives in the literature. In some cases experts had differing opinions on how functions should be

represented. David and Karen used the experts comments to determine which view prevailed in the community.

The main phase of ontology development ended in May 2011 when we removed (merged or obsoleted) 50 of the terms which had been particularly problemetic, while obsoletion of 40 other terms was deferred (detailed below). We also fixed smaller issues with the names, definitions, or placement within the ontology of 78 previously existing terms (19 MF & 59 BP). As of May 2011, we had created 188 new terms (149 MF, 38 BP, and 1 CC), and refined. Since that date, 34 additional new terms have been created (14 MF & 20 BP, not including "regulation of x process by regulation of transcription" terms).

Fate of Previously existing problem terms

fate of term	BP	MF	Total
obsolete	4	27	31
merge (same ontology)	18	1	19
Terms removed	22	28	50
obsoletion deferred	5	35	40
Grand Total	27	63	90

Annotation Issues

Reannotation due to removal of old terms

During the course of the ontology development phase, several issues came up relating to the fate of annotations to terms indicated for removal from the ontology. Minimizing reannotation work and providing appropriate indications to annotation groups were the two highest priorities in determining whether to merge a term into another term or to obsolete it. All of these proposals were written up on individual wiki pages linked from the Proposals section of the Transcription overhaul wiki:

* http://wiki.geneontology.org/index.php/Transcription#Proposals

Particularly within the Biological Process ontology, there were a number of merges of old, poorly defined terms into slightly more general Biological Process terms. There was also the merge of "transcription" into the slightly more specific term "transcription, DNA-dependent", and the correlated merges of the corresponding regulation, positive regulation, and negative regulation terms. This decision was made due to the fact that there were tens of thousands of annotations to the "transcription" terms and examination of AmiGO indicated that the essentially all should have been made to the "transcription, DNA-dependent" terms in the first place (most annotated genes were obviously RNA polymerase II transcription factors). Thus, merging the "transcription" terms into the "transcription, DNA-dependent" terms provided the best option for reannotation.

For numbers of annotations to "transcription" terms, see:

*http://wiki.geneontology.org/index.php/Proposal_for_fate_of_%22transcription%22_and_corresponding_regulation_terms

There were also a number of high-level Molecular Function terms (27) that were defined equivalently to a Biological Process term. As there was precedent for merging a Molecular Function term into the corresponding Biological Process term, we initially proposed to merge these into their Biological Process equivalents. However, there was a counter proposal from Val Wood to obsolete instead so the there would be notice that transfer of the annotation from the MF term to be removed to the equivalent BP term might result in a gene not having any direct MF annotation. We agreed to this and obsoleted these terms and provided suggested reannotation terms, an equivalent term in BP as well as one or more MF terms to consider.

There were also a number of MF terms that were recommended for obsoletion, but for which obsoletion was deferred. These were 13 terms for binding to specific DNA transcription-regulatory elements and 22 terms for binding to specific transcription factors. The argument for obsoleting them was that they represented such a small subset of the possible transcription regulatory sequence elements and transcription factors that they did not serve any useful purpose in providing information about the target binding sites. Instead, the existing information should become part of the annotation process and should be captured in a gaf file for use when a robust system for annotating targets, possibly via column 16, is in place. The counter argument was that we would lose data if we remove them. The agreement was to maintain the status quo, i.e. these terms were not obsoleted and no more of these terms will be added, continuing the practice that had already been in place for several years. With the exception of "promoter binding", which was obsoleted for additional reasons, none of these terms were heavily used in annotation. Some of these are now included in a new obsoletion proposal.

*http://wiki.geneontology.org/index.php/Proposal_to_obsolete_%22promoter_binding%22_and_child_terms#Total_number_of_annotations_per_term
*http://wiki.geneontology.org/index.php/Proposal_to_obsolete_children_of_%22transcription_factor_binding%22#Total_number_of_annotations_per_term

Annotation guidance for annotators

To help annotators understand the upcoming new transcription factor terms, Rama organized an Annotation call focused specifically on transcription on Sept 17th, 2010, shortly after the Bar Harbor meeting. At the first half of this call, Karen went over the basic organization of the new structure in the Molecular Function ontology to represent transcription factors, using the same presentation given at the Bar Harbor. For the second half, she talked about some of the effects on annotation. Once the main ontology development phase was completed, Rama organized a transcription jamboree on July 26th, 2011. For this call, annotators had posted a number of questions on specific papers or genes. As a group, we talked through these examples. Karen also prepared an annotation guide to help annotators understand the new transcription factor activity terms in MF. Perhaps unsurprising considering the number of MF terms that were obsoleted on the grounds that they were essentially identical to BP terms, one of the major issues related to a paradigm shift in what constituted evidence for a MF term; merely knowing that it was involved in transcription without knowing anything about how is sufficient for a BP term, but not for one of the new MF terms.
*http://wiki.geneontology.org/index.php/Transcription_jamboree

Changes in annotation procedure (for some) due to has_part relationships

Implementing the has_part relationship to indicate when a "complex" molecular function, e.g. a "transcription factor activity" has one or more binding activities via the has_part relationship has effectively required changes in annotation practice for some groups. Groups that preferred to annotate only to the most granular MF term and not annotate separately to sub parts of that MF activity have been encouraged to annotate the component binding activities directly, e.g. terms for various specific types of DNA binding, transcription factor binding, RNA polymerase binding, etc. However, it should be noted that this decision requiring use of has_part to indicate component binding activities was made at the Annotation Camp in Geneva, not specifically as part of the transcription overhaul. The transcription factor activity terms in MF just happened to be the first terms implemented in this way.
*http://gocwiki.geneontology.org/index.php/Talk:2010_GO_camp_Meeting_Agenda#Binding_continued

Fundamental disconnect between scope of ontology and annotation projects

It seemed like an obvious choice to do an annotation project using the new terms just created by the transcription ontology overhaul. However, the scope of the ontology project was specifically limited to the long-standing issues in the MF ontology and a few major issues at the top of the BP ontology. In contrast, the scope of the heart transcription factor annotation and transcription factor reannotation (due to removed terms), projects were more comprehensive with respect to the terms used for annotation. Many of the transcription factors involved in heart development, which were the targeted focus of that project are represented by a single term in the MF ontology (one representing sequence-specific DNA binding transcription factors for RNA polymerase II). For reannotation of genes for which the annotated terms had been removed, at least in groups that comprehensively reannotated, there was usage of a slightly larger set of the new MF transcription factor activity terms.

However, a significant amount of the annotation effort focused on BP. Thus, the majority of requests for new terms were requests for new "regulation of x process by regulation of transcription" terms. Terms like this were not within the scope of the transcription overhaul. This expansion in scope opened up subsequent ontology-developmentissues with respect to compound terms that describe essentially the regulation of some process A by means of regulation the transcription of gene products involved in process A itself. This issue is still under discussion by the ontology editor's group (see minutes of recent ontology development calls). In particular we foresee the expansion of these terms to cover almost every process in the BP area of GO. It was not straightforward to create a TermGenie template to cover these terms due to dual is_a parentage. It is still not clear whether it would be better to handle these annotations via a more sophisticated annotation paradigm that allows linking together of separated statements. For example 'regulation of glucose metabolism by regulation of transcription from RNA polymerase II promoter" might be handled by creating a complex annotation to 'regualtion of transcription from RNA polymerase II promoter' and 'regulation of glucose metabolism'. Many of these requests were eventually added to the ontology by hand to alleviate the backlog that was being generated by annotator requests. This decision was also discussed at an ontology developer's call.

Summary Comments

- Having a team, rather than a single developer, to be able to discuss ideas was incredibly beneficial. In this case, both ontology developers were also annotators capable of bringing that perspective to the table. Frequent communication was essential.

- When decisions affecting ontology development are made at a meeting, there should be more effort to make sure that ongoing ontology development projects are informed of decisions that may impact their development work.

- For any project fixing long standing issues within GO or changing fundamental aspects in the way terms are represented, the expectation should be that it will require significant input of time, both by ontology developers and by annotators, and may require annotators to adjust their annotation practices based on the new structure.

- If, after a significant investment of time researching possible solutions to a problem in the ontology, ontology developers recommend removal of terms that some annotators may be fond of, we need to have a rational cost benefit analysis of the benefit of keeping those terms and annotations versus the cost of leaving problem terms in the ontology, potentially increasing the the time needed to fix both the ontology and annotations in the future.

- We took advantage of biological subject expertise within the GOC. This was very productive as these people already understood GO and were motivated to improve it.

- Consulting outside experts when necessary was enlightening and productive in resolving tricky issues. We found it effective to just contact appropriate people as needed rather than having an ongoing panel.

- Providing opportunities for annotators to learn about the new developments and talk about specific examples seemed very helpful in reannotation.

- If the ontology development and annotation projects are sequential, it would be good to plan to have time dedicated by an ontology developer for continued work in conjunction with the annotators.

- We need to learn how to handle the circumstance when an annotation effort supporting ontology development goes far beyond the scope of the work done for the ontology development project per se. In this example, annotation of molecules that play a role in transcription with respect to other processes went beyond the molecular function aspects of the ontology development work and opened up significant issues in ontology development. How should we handle this?