A review of the Gene Ontology: past developments, present roles, and future possibilities.

Introduction

The Gene Ontology (GO) was one of the most important first steps towards solving one of the greatest computational problems of biology: unambiguously representing all biological knowledge in a computationally tractable way. The original practice of representing most essential biological facts using qualitative description in our infinitely expressive human tongue is unfortunately incomprehensible by computers. The challenge is to create this systemized representation while maintaining as much of the subtle truth contained within those descriptions as possible. This paper covers how GO has approached this overwhelming challenge by starting small and carefully expanding its focus, as well the challenges and solutions it has encountered along the way and the ones it might yet face.

I. The Foundation of GO

A paradigm shift in biological research emerged in the late 1990s as the field, classically based on description, became increasingly data-driven. Biological databases had steadily grown in quantity and use since the 1970s, when DNA sequencing was invented[1] and the Protein Data Bank (PDB) was founded. By the late nineties there were many more databases, such as SwissProt for annotated protein sequences, and Flybase, AceDB, and SGD for individual model organisms' genomes. These databases thrived within their respective scientific communities, but had zero connection between them[2]. As the untapped potential of interconnecting biological knowledge on a global scale was being recognized, the large scale functional analyses enabled by the invention of DNA microarrays[3] further demanded a soundly thought-out integration of genomic databases.

Representatives from the yeast, mouse, and fly model organism databases founded the Gene Ontology Consortium in 1998 to collaborate on a methodology of integrating the information contained within their databases[4]. The original intent of the group was just to create a set of standardized vocabularies they could share, but they quickly realized the great value that their combined data and a globally formalized semantic schema could have for the rest of the scientific community[5]. Corresponding genes were not consistently annotated from database to database[6], and so the Gene Ontology was designed to make these free-text based annotations tractable and consistent[7]. Both the structure of how the annotation was formed and the terminology used within the annotation would be standardized. GO was described as a "structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism" in a 2000 paper boldly titled "Gene Ontology: tool for the unification of biology"[8].

II. The Construction of GO

Due to the immediate demand for an integrated and standardized gene annotation database and the many pitfalls potentially involved, the curators wisely chose to simplify things as much as possible[4]. Their efforts were focused on defining the words needed describe particular features of biology— they wrote that they were "*aware that this is an incomplete solution, but firmly believe that it is a necessary first step"* and that the sets of terms themselves would be *"immediately useful"*[5]. So despite its name, the GO is not an 'ontology' as classically defined by computer scientists and philosophers, but a 'controlled vocabulary'[9]. Significantly, the two components of GO, the terms composing the vocabulary and the annotations of the terms to genes and gene products (which will hereafter be referred to simply as genes), were independently developed from the start[5].

GO terms are grouped into one of three separate vocabularies chosen to represent sets of information shared by all life and fundamental to describing a gene[5]. These three vocabularies are (1) cellular component (where is it), (2) molecular function (what it does there), and (3) biological process (how it does it)[10]. GO structures each domain as a "directed acyclic graph" (DAG, see Fig. 1[11]), a type of tree where a term can have zero or more children and one or more parents[12]. This permits a hierarchy while allowing terms to be defined by multiple types of categories— allowing, for example, conveyance that "endoribonuclease activity" is a subset of both "endonuclease activity" and "ribonuclease activity"[10, 12]. Terms can also be linked as synonyms. A GO term entry consists of its name, unique ID, definition with cited sources, and a reference to which of the three domains it belongs[8].

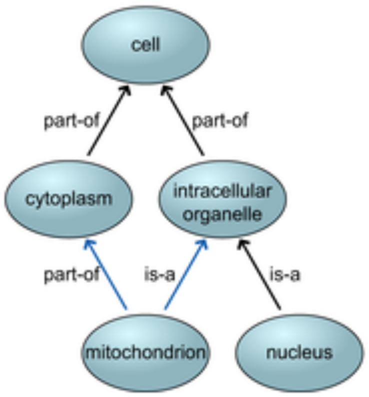

Figure 1. - Directed Acyclic Graph

At GO's outset, parent-child relationships were described by only one of two definitions: "is_a" and "part_of"[4]. The 'is_a' relationship indicates that one term is a subclass of the other;

for example, a "neuronal cell body" is_a "cell body" is_a "cell part"[10]. The relationship is transitive, so "neuronal cell body" is_a "cell part" also. The "part_of" relationship is more complex: in the 'cellular component' domain it means 'is physically part of' and in the 'biological process' domain it means 'is a subprocess of'[12]. A term many only be part_of one parent, but may have several 'is_a' parents.

Curation of GO term annotations to genes has been user-driven from the start[6]. An annotation must indicate its source, usually a literature reference, database, or computational analysis, as well as an 'evidence code' indicating what type of experiment supports the annotation[13]. At first relationships were manually annotated as needed, usually just a few at a time. When conflicts arose, such as with the logic of a DAG's structure, the consortium would discuss the problem and deal with it on a case-by-case basis[14]. This strategy would become less feasible as the size and complexity of the GO grew, and automated methods for contributing and quality-checking annotations would be required[5].

III. Mixed Early Reception

After GO was introduced, it quickly grew in popularity but also faced its fair share of outspoken critics in the scientific community. These critics can be generally classified into two groups. One group, characterized as more classical biologists that hadn't yet comprehended the paradigm shift towards data-driven biological research, basically saw the whole GO effort as naïvely misguided and disagreed that it would have any value. The second group, characterized as having more expertise in fields like informational science, computer science, and philosophy and having less expertise in biology, felt that the GO curators had the right intentions but made too many logical simplifications and compromises that rendered the whole enterprise useless[4].

The best example of the first group is found in Sydney Brenner's 2002 journal comment "Ontology recapitulates philology"[15]. Brenner, a highly respected biologist and Nobel laureate, broadly critiques the whole idea of GO with great wit but also an apparent blind spot for GO's valuable purpose. He claims somewhat vaguely that "*the network we should be interested in is not the network of names but the network of the objects themselves*", by which he seems to mean their sequences, phenotypes, etc. The quickly published response, aptly entitled "Ontologies for programs, not people"[16] defends GO on several points. Firstly, the term names are hardly the essential elements of GO, they are just reflections of language already used in the literature, necessary to convey the truly important elements: the relationships. Secondly, GO is accessed primarily by computer programs, not directly by the user. Even if GO is far from a perfect representation of reality, it enables many bioinformatic approaches that can extract probalistic truths from the information within GO. The final fact is that the exponential growth of published data and gradual integration of sub-disciplines demand some computerized organization, and the GO is certainly at least a step in the right direction.

Most other critiques came from scientists whose expected visions for GO were unaligned with or too ambitious for the GO consortium's initial vision. Many criticisms seemed to stem from the simple fact that GO is not an 'ontology' as traditionally defined, despite its name, as well as that its authors focused their energy on its practical use and biological meaning rather than its theory, logic, or code. A traditional ontology is supposed to have a formal specification and definition of its categories and relations, which GO lacks[17]. A 2003 paper "The Ontology of the Gene Ontology"[9] states that the authors of GO faced a "*trade-off between (1) formal and ontological coherence, stability and scalability, and (2) the speedy population of GO with biological concepts"* and posits that too little attention was given to the former. One example of a lack of logical and ontological rigor that was pointed out was that the DAG representation was not formalized at first[17]. Another source of issues was the sometimes fuzzy distinction between the 'biological process' and 'molecular function' domains—having a process term like 'transport' and a function term like 'transporter' can lead to confusion[18].

Some biologists expected the GO to let them describe specific events under specific conditions[4]. GO was critiqued for its lack of logical expressive power due to its limited set of ways to describe a relationship. Furthermore, of the two existing relational terms, the *part_of* relation had been used inconsistently and created confusion[17]. Some of these criticisms were based on expectations too lofty, but many others have since been addressed, such as the paucity of ways to describe a relationship.

IV. The Improvement of GO

The current version of GO has over 30,000 terms and 50,000 relationships[19]. Since its inception, GO has steadily grown in every aspect: the quantity of terms, annotations, relationship types, organisms covered, third-party tools, citations (see Fig. 2[11]), web-site use, and the quality of logical rules, automated checking, educational resources, and resources for various special interests and subdisciplines. Like the biological knowledge it seeks to reflect, GO is ever-changing, existing relationships and terms are refined and reorganized as the current state of

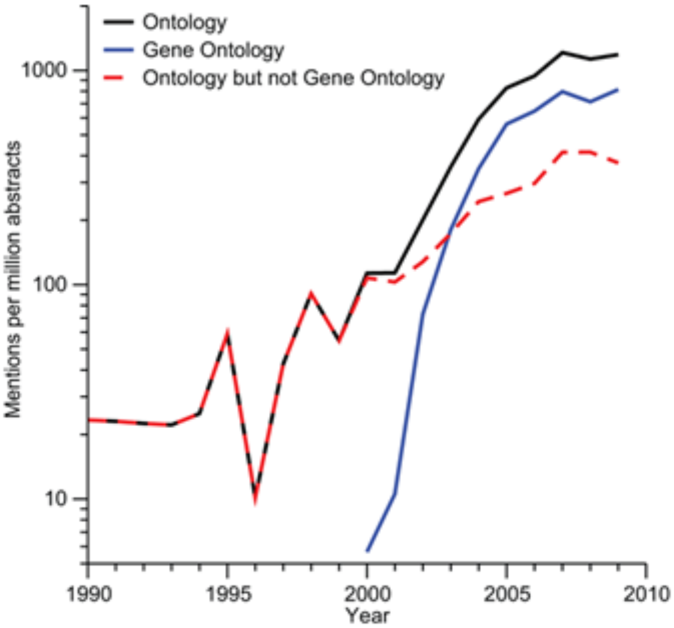

Figure 2 - Growth of mentions per million PubMed abstracts

knowledge advances[20]. GO curators, annotators of model organism databases and other interested biologists propose changes through an online tracking system, which are reviewed by the GO Editorial Office and usually result in some change to the database. The logical theory underpinning GO has strengthened, sometimes resulting in large-scale changes, such as in 2003 when all molecular function term names had the word 'activity' appended to them[9].

While the essential foundation of GO is that it generalizes and integrates among subdisciplines, many steps have been taken to accommodate biologists who only need be concerned with specific sections of the GO database, notably through the concept of 'GO slims'. A GO slim is a narrowed-down version of the ontology that contains only a subset of the total terms[19]. They are created by users as needed are either specific to a species or some particular area of biology, and there is a whole archive of them publically available[19], for example, one for 'Honey Bee ESTs'.

The Sequence Ontology (SO) was a sister project to GO established in 2003 to '*promote the standardization of sequence annotation among different organisms*'[21]. It provides the tools and terms for describing actual DNA, RNA, and protein sequences among different organisms. This created a fourth non-overlapping domain of knowledge encompassed by GO, that of sequence features[12].

One significant area of improvement was with the types of relationships that could link terms[22]. Relationship terms 'regulates', 'positively_regulates', and 'negatively_regulates' were added in the past two years to enable GO to distinguish between when a process affects another's manifestation but does not play a direct role in it. Also, 'has_part' was added to give the parental complement to 'part_of'. The half dozen relationship types now allow for some fairly elaborate logical reasoning on GO—for example, if A 'is_a' B and B 'regulates' C and D is 'part_of' C, then we can conclude that A 'regulates' D. Also, links are now allowed between the molecular function (MF) and biological process (BP) domains—an MF term may be part_of a BP, and both BP and MF can have a 'regulates' relationship.

As the GO authors originally anticipated in 2001: *"[It will] be increasingly difficult to maintain the semantic consistency we desire without software tools that perform consistency checks and controlled updates"*[5]. Thus many new quality checks, both automated and manually conducted, have been introduced to GO over the last decade[22]. For example, a biological validation has been undertaken to compare annotations of overlapping sets of genes with those that would be expected to be mutually exclusive, revealing errors that either derive from the ontology's structure or the annotation. Another example is a check to make sure a given species' gene isn't annotated to a process that species is incapable of, like *homo sapiens* and photosynthesis.

GO has also improved itself through many more practical ways. The website's interface has been redone several times and is now simple, efficient, and thoroughly cross-linked with other databases[19]. Educational outreach to the researchers and tool-developers that are on the user-end has been enhanced. For a while there were 'user meetings', open to non-consortium members where practices are reviewed and discussed and education about GO is spread[20], but these types of efforts have since been distributed to internet-based strategies. There is now an excellent GO wiki and helpful online community, and the process for accepting user input and suggestions has become more accessible[19].

V. GO Usage Today

Today GO is used for many reasons, as described in the GO Usage FAQ[19]:

- integrating proteomic information from different organisms;
- assigning functions to protein domains;
- finding functional similarities in genes that are overexpressed or underexpressed in diseases and as we age;
- predicting the likelihood that a particular gene is involved in diseases that haven't yet been mapped to specific genes;
- analyzing groups of genes that are co-expressed during development;
- developing automated ways of deriving information about gene function from the literature;
- verifying models of genetic, metabolic and product interaction networks.

One of the most common and important uses of GO is to characterize results from high-throughput studies. GO is one of the few ontologies used to describe such large datasets and succeed in revealing trends that may have been missed otherwise[11]. Usually the user has a large set of gene expression data, and they want to find a cellular component, biological process, or molecular function that is over or under-represented, if some new function can be inferred from the terms' characteristics, or how genes are distributed between some set of biological categories[23].

'Functional profiling' is when one tries to find differentially expressed categories of genes between sets derived from different conditions, such as wildtype versus knock-out[23]. The challenge is to differentiate between which terms are truly 'enriched' or 'un-enriched' and terms that appear to be so by chance, and overlooking this sometimes enables researchers to find the results they wanted right out of thin air. Statistical corrections can help reduce this problem, but a more powerful solution is to have multiple 'replicates' for one's test sets, such as tissues from 4 wildtype mice and 4 knockout mice, all raised in as similar conditions as possible. Of course, multiple test sets may create more problems than they solve, and a multiple comparison correction such as the Bonferroni must be applied[24].

A wide variety of functional profiling tools exist. The fact that entering the same data set into many different such tools results result in p-values varying by several orders of magnitude for some GO terms[25] is indicative of both the unfortunate inexactness of using GO for research and the importance of considering what tool one uses when there are multiple options. The best strategy is always of course to try as many available tools as possible and compare and contrast the results.

'Functional categorization' of a set of genes into subcategories based on shared high-level GO terms is another common application. This is an effective way to efficiently convey a broad characterization of a specific set of genes, differential expression patterns, or a particular genome[23], such as in this graph from a paper presenting the draft sequence of the rice genome[26] (Fig. 3). The GO

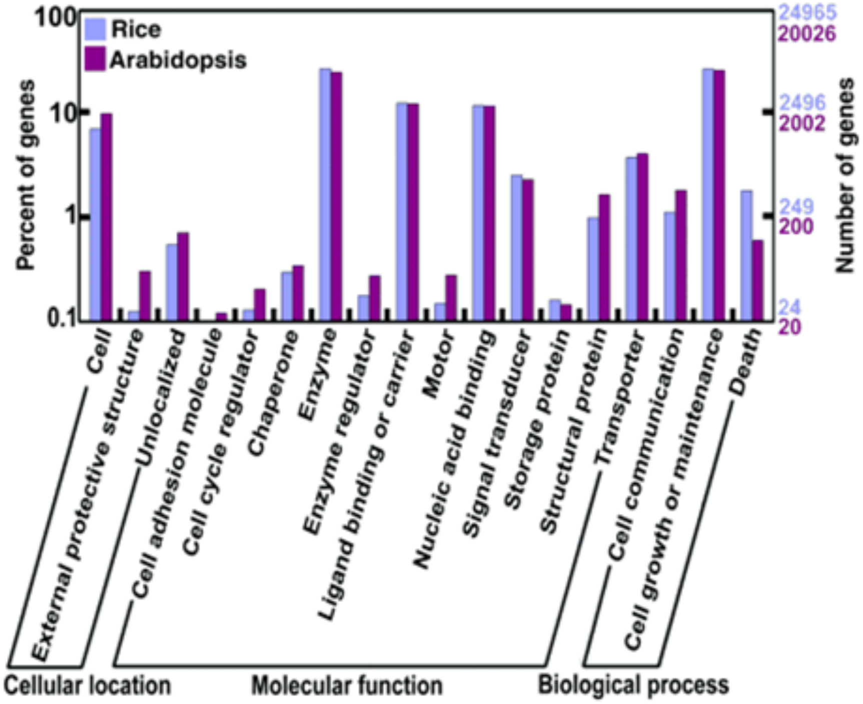

Figure 3 - Functional categorization of yeast genome genes

website offers an algorithm 'map2slim'[27] that maps annotations to a GO slim, which should be carefully chosen by the user as appropriate for the given organism and context of the annotations.

Predicting gene function is performed in several ways; typically a variation of the 'guilty by association' strategy. Enrichment of terms is calculated for a group of partially-characterized genes, and the uncharacterized genes are assumed to be involved with the enriched terms as well[23]. An understanding of annotation practices and the underlying biology is necessary to reach reasonable conclusions. For example, inferences based on correlated gene expression are more appropriate for biological processes than cellular components.

Two key aspects of annotation practice commonly overlooked are the 'NOT' qualifier and evidence codes[4]. Sometimes annotations are created with the 'NOT' qualifier to specifically say that a gene is specifically not associated with a term one might reasonably expect it to be[19]. Obviously, overlooking this detail can produce results that are blatantly incorrect. Evidence codes are very important because they indicate how much one should trust a given annotation. There are two main evidence code categories: experimental and computational. It is sensible to give more credence to experimentally verified data as it actually reflects some real-world event, and assume that computationally inferred data will give one more false positives, especially if it was not manually curated—and note that 95% of all annotations were computationally inferred and automatically generated (these have the 'IEA' code) [23]. Finally, the code 'ND', 'no biological data available' is of special note. It indicates the 'known unknowns', meaning the curator performed an exhaustive literature and found nothing.

VI. The Future of GO

There is no reason not to expect the continuation of the trends of growth and development GO has experienced over the last decade, as well as new ways of GO evolving. Each year GO partners with a few new databases and projects, each bringing with it unexpected insights and challenges. GO has always been an unprecedentedly useful tool for biologists who have interest in relatively general genetic patterns and/or those that use the most popular model organisms. However, there remain many groups of biologists with very specific interests that GO could potentially spread its attention to. One of the most recent collaborations, for example, was with the Plant Associated Microbe GO (PAMGO), which resulted in the creation of over 700 new GO terms, mainly dealing with '*biological processes common to diverse plant- and animal-associated microbes'*[28].

Efforts like the partnership with PAMGO are important for the continued deepening of the biological knowledge already formalized by GO, but some of the most exciting developments will result from the ability to ask new types of questions. The Open Biological and Biomedical Ontologies (OBO) project is a collaborative effort between GO and other ontologies to assist in their sharing principles and practice[29]. Collaborations with databases that store types of information fundamentally different than that currently well-covered by GO and its connected databases all offer their own exciting possibilities. Plans for integration include that with ontologies for phenotypes, anatomy, cell types, diseases, and signaling pathways[4]. An example of an area that deserves more attention in the future is the interaction of species with their environments, such as similarities in lifestyle and habitat[11].

Whenever the structure of GO is expanded, it creates new ways of expressing facts and queries, but these take time to become established in research. For instance, functional profiling a group of genes to find those that correlate is now quite popular and easy thanks to the many available tools. But with the introduction of the new relationship types, there is now more subtle truths available to a researcher if he or she is willing to put forth the extra effort to parse them out. The GO Consortium hopes that people will use these newest features to ask more 'hypothetical questions' of GO: "*For example, a user could now ask what gene products might be involved in regulating a specific metabolic process if they know a regulatory process that controls the metabolic process and they know the types of molecular functions that play roles in the regulatory process*."[22]

Making changes to a database like GO's can be very challenging due to the high degree of connectivity between it and the many external entities that are dependent on its current state. For example, there is a growing concern that the columns on the annotation form are getting overloaded in use and need to be redone[4]. However, this has yet been concluded to be more trouble than it is worth, because of all the third-party tools that would need to be changed as well. This drawback of dependently developed projects is a significant and representative problem. Moving forward, changes should be made with a perspective that stretches many years ahead, and expects and accommodates potential future changes as well as possible.

One exciting trend will come from the potential for using GO to help make new discoveries. Only a few examples of realizing this potential have occurred thus far[11]. A rare early example was the Genes2Disease method developed in 2002 which predicts disease-related genes by correlating GO molecular functions of those genes with disease and phenotype MeSH terms[30]. Two papers have been published recently that made new discoveries using ontologies: one links animal models with human diseases by comparing phenotypes and anatomy across species[31], and another finds undiscovered drug targets by systematically comparing side effects[32]. Both studies used the same general strategy: link abstract concepts, be they diseases, drugs, or animal models, based on their associated phenotypes[11]. In a promising 2009 paper, Mungall et al. describes a method for using cross-product mappings of multiple OBO databases to perform cross-ontology queries[33].

Taking a long-term, big-picture view, one could envision the ramifications of GO and its interconnected databases expanding to encompass more and more of the global biological knowledge. It's possible the process of sharing a finding would involve entering all the information one would today convey through the language of the published scientific paper into a computer program, which automatically parses it out and integrates it into the appropriate databases. The ideal endpoint of this would be when our ability to organize and integrate the data we have once again exceeds our ability to gather it, which has exploded thanks to high throughput technologies. Hopefully this will allow the biological community to concentrate its efforts on the most interesting questions, the ones only a human mind can comprehend.

References

1. Sanger, F. and A.R. Coulson, *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* Journal of Molecular Biology, 1975. 94(3): p. 441-446.
2. Lewis, S., *Gene Ontology: looking backwards and forwards.* Genome Biology, 2004. 6(1): p. 103.
3. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science (New York, N.Y.), 1995. 270(5235): p. 467-470.
4. Cherry, M., *Interview*. 2010.
5. Consortium, T.G.O., *Creating the Gene Ontology Resource: Design and Implementation.* Genome Research, 2001. 11(8): p. 1425-1433.
6. Schuurman, N. and A. Leszczynski, *Ontologies for Bioinformatics.* Bioinformatics and Biology Insights, 2008. 2008(BBI-2-Schuurman-et-al): p. 187.
7. Lord, P.W., et al., *Semantic similarity measures as tools for exploring the gene ontology.* Pac Symp Biocomput, 2003: p. 601-612.
8. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nat Genet, 2000. 25(1): p. 25-29.
9. Smith, B., J. Williams, and S. Kremer, *The Ontology of the Gene Ontology.* 2003: p. 609-613.
10. *AmiGO - official online tool-set for searching GO*. 2010 December 4th, 2010]; Available from: http://amigo.geneontology.org.

11. Jensen, L.J. and P. Bork, *Ontologies in Quantitative Biology: A Basis for Comparison, Integration, and Discovery.* PLoS Biol, 2010. 8(5): p. e1000374.
12. Harris, M.A., et al., *The Gene Ontology project*. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 2004: John Wiley & Sons, Ltd.
13. Consortium, G.O., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Research, 2004. 32(suppl 1): p. D258-D261.
14. Hill, D., et al., *Extension and Integration of the Gene Ontology (GO): Combining GO Vocabularies With External Vocabularies.* Genome Res., 2002. 12(12): p. 1982-1991.
15. Brenner, S., *Life sentences: Ontology recapitulates philology.* Genome Biology, 2002. 3(4): p. comment1006.1 - comment1006.2.
16. Hunter, L., *Ontologies for programs, not people.* Genome Biology, 2002. 3(6): p. interactions1002.1 - interactions1002.2.
17. Poli, R.H., Michael; Kameas, Achilles *Theory and Applications of Ontology: Computer Applications*. 1 ed. 2010: Springer. 400.
18. Kumar, B.S.A., *On Controlled Vocabularies in Bioinformatics: A Case Study in the Gene Ontology.* BIOSILICO: Drug Discovery Today, 2004. 2: p. 246–252.
19. *The Gene Ontology Website*. 2010; Available from: http://www.geneontology.org.
20. Consortium, G.O., *The Gene Ontology (GO) project in 2006.* Nucleic Acids Research. 34(suppl 1): p. D322-D326.
21. Eilbeck, K., et al., *The Sequence Ontology: a tool for the unification of genome annotations.* Genome Biology, 2005. 6(5): p. R44.
22. Consortium, T.G.O., *The Gene Ontology in 2010: extensions and refinements.* Nucleic Acids Research, 2010. 38(suppl 1): p. D331-D335.
23. Yon Rhee, S., et al., *Use and misuse of the gene ontology annotations.* Nat Rev Genet, 2008. 9(7): p. 509-515.
24. Farcomeni, A., *A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion.* Statistical Methods in Medical Research.
25. Khatri, P. and S. Drăghici, *Ontological analysis of gene expression data: current tools, limitations, and open problems.* Bioinformatics, 2005. 21(18): p. 3587-3595.
26. Yu, J., et al., *A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica).* Science, 2002. 296(5565): p. 79-92.
27. Mungall, C. *CPAN - map2slim*. 2010; Available from: http://search.cpan.org/~cmungall/go-perl/scripts/map2slim.
28. Torto-Alalibo, T., C. Collmer, and M. Gwinn-Giglio, *The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions.* BMC Microbiology, 2009. 9(Suppl 1): p. S1.
29. *OBO - Open Biological and Biomedical Ontologies Website*. 2010; Available from: http://www.obofoundry.org/.
30. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. 31(3): p. 316-319.
31. Washington, N.L., et al., *Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation.* PLoS Biol, 2009. 7(11): p. e1000247.
32. Campillos, M., et al., *Drug Target Identification Using Side-Effect Similarity.* Science, 2008. 321(5886): p. 263-266.
33. Mungall, C., et al., *Cross-Product Extensions of the Gene Ontology.* Nature Precedings, 2009(713).