Identifying Gene Ontology Areas for Automated Enrichment

Catia Pesquita and Francisco Couto

LaSIGE, Universidade de Lisboa
Campo Grande, Lisboa, Portugal

Abstract. Biomedical ontologies provide a commonly accepted scheme for the characterization of biological concepts that enable knowledge sharing and integration. Updating and maintaining an ontology requires highly specialized experts and is very time-consuming given the amount of literature that has to be analyzed and the difficulty in reaching consensus.
This paper outlines a proposal for the development of automated processes for the enrichment of the Gene Ontology (GO) that will use text mining techniques and ontology alignment techniques to extract new terms and relations. We also identify the areas of GO whose level of detail is too low to answer the community's needs at large. We have found that although GO's content is well suited to the manual annotations, revealing the coordination between GO developers and GO annotators, there are 17 areas that would benefit from enrichment to support electronic annotation efforts.
With this work we hope to provide biomedical researchers with an extended version of GO that can be used 'as is' or by GO developers as a starting point to enrich GO.

Key words: Biomedical ontologies, ontology enrichment, text mining, ontology alignment

1 Background and Research Problem

In recent years, biomedical research has generated an enormous amount of data that is spread across a large number of repositories, which are often publicly available on the Web. With this, finding the relevant sources and retrieving the relevant information has become a non trivial task. One important breakthrough in this area was the development of biomedical ontologies. In the bioinformatics domain, the term ontology can have a wide range of meanings, from controlled vocabularies, taxonomies, thesaurus and frame-based systems to rich logical axioms encapsulating our knowledge [3]. Briefly, an ontology should contain formal explicit descriptions of the concepts in a given domain, which should be organized and structured according to the relationships between them.
Developing a domain ontology is a very complex task, that involves high expertise both over the domain to model and in knowledge engineering. Developing

an ontology for the biomedical domain represents an even more interesting challenge given the speed at which biomedical knowledge is growing, particularly since the advent of high throughput techniques. This means that a biomedical ontology can never be considered complete, and that the effort to maintain these ontologies is very heavy.
In order to alleviate this problem, ontology enrichment techniques can be employed. These are automated processes that identify new candidate concepts to add to the ontology or new relations to be instantiated. Ontology enrichment is built upon the techniques used for automated or semi-automated ontology construction, and brings together several disciplines, including natural language processing, data and text mining, machine learning and clustering.

The flagship of biomedical ontologies is the Gene Ontology (GO) [8]. It is currently the most successful case of ontology application in bioinformatics [1], and provides an ontology for functional annotation of gene-products in a cellular context, capable of dealing with the semantic heterogeneity of gene product annotations in other databases. GO comprises three aspects (or GO types):

- Molecular Function: processes at the molecular level.
- Biological Process: assemblies of various molecular functions.
- Cellular Component: cellular locations and macromolecular complexes.

It is structured as a directed acyclic graph (DAG), where each node in the graph is a natural- language term describing a biological concept within GO's domain; and each edge represents a relationship between terms, that can fall within five types: is_a, part_of, regulates, positively_regulates, negatively_regulates. It is important to stress that GO only represents classes (concepts describing functional aspects of gene products) and never the real instances (gene products themselves).
The Gene Ontology was developed by the GO Consortium, initially a collaboration between three model organism databases (FlyBase, Saccharomyces Genome Database (SGD) and Mouse Genome Informatics (MGI)), to address the need for a common and consistent vocabulary to annotate gene-products of different databases. Nowadays the Gene Ontology aims at being species independent and the GO Consortium has grown to fifteen members which cooperate in maintaining and updating GO. It has grown from about 3500 terms in 1998, covering three databases to currently over 20,000 terms spanning about 20 databases.
The primary functionality of GO, the annotation of gene products, is largely achieved by the GOA project [5], which provides GO term annotations for gene products present in UniProt and other major databases.
GO is a handcrafted ontology, where members of the GO consortium group contribute to its updates and revisions. There are about 100 contributors to GO spread across the several GO Consortium and GO Associates members, and they are expected to contribute regularly towards the content of GO. Since GO covers a broad range of biological areas, GO has setup interest groups to discuss the areas within the ontology that are likely to require extensive additions or revisions. These groups roughly correspond to high-level terms: cardiovascular, developmental biology, electron transport, farm animals, immunology, metabolism, neu-

robiology, pathogens and pathogenesis, protein kinases, response to drug, and transport. Other GO users can also contribute by suggesting new terms via Sourceforge.net, however the majority of content requests are made by GO team members (see Table 1).

	people	total requests	request/person
GO members	53	2545	48.02
External users	46	337	7.33

Table 1. Summary of new GO term requests on Sourceforge.net

We believe that it would be of great importance to develop methods that could help the GO team to develop GO in a more efficient manner, and that ontology enrichment processes can play a major role in this. There are number of resources that can be capitalized by ontology enrichment techniques to boast GO extension, namely the large amount of publicly available biomedical literature and the many biomedical ontologies and terminologies.

2 Related Work

The automated enrichment of biomedical ontologies is still in its early steps, with few works in existence: [15] propose a method based on verb patterns to enrich a molecular interaction knowledge base;[10] propose a method to expand GO outside its 3 areas by combining two orthogonal vocabularies; and [13] uses the syntactic relations between existing GO terms to propose new ones.
However, there are many efforts for automated ontology learning outside the biomedical domain: [12] uses algebraic extraction techniques to convert a dictionary into a graph structure. [6] uses word usage statistics from a text corpus constructed through mining the web, and [17] uses a text mining approach to generate groups of related terms to propose to the ontology engineers. [9] uses lexico-syntactic pattern matching to learn new relationships between concepts in an ontology.Several clustering methods have also been developed for learning ontologies from text corpora [7], [2],[16]. [18] uses name matching methods based on machine learning to identify new concepts while [11] uses Formal Concept Analysis to derive a concept hierarchy from syntactic dependecies extracted from a text corpus.

3 Research Methodology

This work is composed of five tasks:

1. Identifying areas of GO where enrichment can be beneficial - GO 'hotspots'

2. Developing text mining methods to extract new terms and relations from publicly available texts
3. Developing ontology alignment methods that will enable the reuse of other ontologies by GO
4. Integrating the new terms and relations into GO's structure
5. Evaluating the results of the enrichment

Figure 1 summarizes the articulation between these tasks.

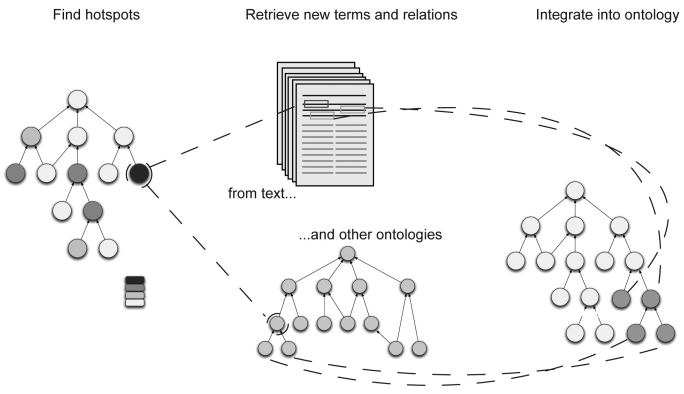

Fig. 1. Workflow for the automated ontology enrichment project

The first task is already underway and we present some results pertaining to it in the following section. The other tasks are still at a planning stage, so we present an overview of each.

3.1 Identifying GO 'hotspots'

The main idea behind this task is that the areas of GO that would benefit the most from automated enrichment, would be the ones that are lagging behind in size but still boast a significant usage for annotation. We define GO 'hotspots' as areas where the current level of detail is not answering the community's needs. To this end, we consider as distinct GO areas the most specific terms within GOSlim generic [1], and identify which ones are promising spots for enrichment. This will be followed by a more in depth analysis that will focus on single terms who have a pattern of annotation that indicates a need for more specificity.
Here, we have analyzed the pattern of annotation of the 87 leaf terms of GOSlim generic across five versions of GO distributed over a period of two years, and also analyzed the evolution of those areas regarding number of terms.

[1] GOSlims are subsets of GO, that only include high-level terms and aim at summarizing GO. Each leaf term, the most specific terms in GOSlim, is a representative of all its children terms and their annotations.

3.2 Text Mining

After identifying the areas to enrich, we will apply text mining techniques to two text sets: one will be based on the automated retrieval of relevant abstracts from PubMED, while the other will be a smaller corpus composed of manually selected full texts. The techniques to address new terms and relations extraction will have to address several issues, including the compositionality of GO terms (e.g. 'transport' and 'ion transport'), the high degree of synonymy and homonymy in biomedical vocabulary, distinguishing between ontology concepts and instances and distinguishing between the different kinds of relations.

3.3 Ontology Alignment

It is also possible to propose new terms to GO by aligning GO with other relevant ontologies (such as the Signal Ontology, the Cell Ontology, ChEBI) and integrating them with GO. We will combine several ontology alignment strategies that exploit distinct sources of information: labels and descriptions of the terms, domain knowledge extracted from literature; structural information, particularly the different types of relationships; and annotations from the GOA database to deduce similarities between concepts based on the instances classified into them. We will also investigate the application of previously developed semantic similarity measures to this task [14].

3.4 Ontology Enrichment

To be able to propose valid new terms to GO, the results of the first two tasks need to be checked for their consistency. In the case of terms derived from ontology alignment, we will have to check for conflicting subsumption relationships, which can be particularly relevant since GO is organized as a DAG (directed acyclic graphs), so a term can have multiple parents. Also, GO has different types of relations, and not all of them are transitive over each other.
The next step is to organize the new terms in a hierarchy, in order to reflect their degree of specificity. Both clustering techniques and natural language processing can be employed to this end, since we will need to find the relative specificity of each term in relation to the others. We will also take advantage of the alignments to propose improvements to GO's descriptions of terms, by combining the descriptions of both aligned concepts into a more complete description.

3.5 Validation

We will validate these extensions by running the enrichment method on older versions of GO, and then comparing the extended version to the most recent version. This will allow us to measure the precision of our approach, by verifying if any of the terms we defined for the older version of GO were included by GOs developers in a more recent version.

4 Results and Discussion

The results presented here are preliminary and concern only the first task, the identification of the areas of GO that would benefit the most from automated enrichment. To identify them we have calculated for the 87 leaf terms of GOSlim generic the ratio between the annotations made to that GOSlim term and the number of terms that it represents (number of children). We have computed this ratio for five versions of GO spanning two years. To distinguish between manual annotations and computationally derived annotations, we have calculated two different ratios for each version, one considering just the annotations that are made by curators, and another considering all annotations present in GOA. For these two scenarios we have identified 17 'hotspots' that denote increased annotation activity that is not accompanied by an extension of that GO branch.We considered a GOSlim term to be a 'hotspot' if at any given time a 1.5 fold increase in the ratio of annotations per child was observed, that was not subsequently decreased. Figures 2 and 4 show the distributions of the annotation ratios for these terms in each scenarios.

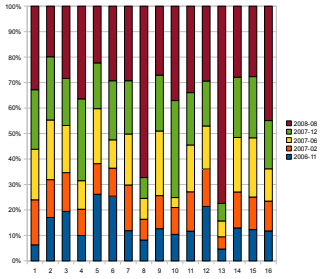

Fig. 2. Distribution of the annotations per child ratio for the 16 'hotspots' found using all annotations. 1) reproduction 2)generation of precursor metabolites and energy; 3) DNA metabolic process 4)cell recognition 5)cell death 6)embryonic development 7)cellular homeostasis 8)cytoplasmic chromosome 9)cell wall 10)lipid particle 11)cilium 12)ion channel activity 13)electron carrier activity 14)antioxidant activity 15)oxygen binding 16)chaperone regulator activity

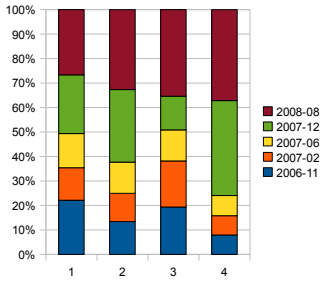

Fig. 3. Distribution of the annotations per child ratio for the 4 'hotspots' found using manual annotations. 1) reproduction 2)embryonic development 3)viral reproduction 4)lipid particle

It is interesting to note that some of these 'hotspot' overlap GO's Interest groups (e.g. embryonic development, viral reproduction, electron carrier activity, generation of precursor metabolites and energy). This is a good indicator that we are in fact identifiying areas of interest. It is also noteworthy that the

number of identified 'hotspots' when using manual annotations is very low, only four, when compared to the number of 'hotspots' identified when considering all annotations, 16. There is considerable overlap between these two sets, with only one term being identified exclusively by the manual annotations approach, 'viral reproduction'.
The low number of manually originated 'hotspots' may be a reflection of a good articulation between GO content development and GO manual curation, which can mean that many GO terms are created when GO curators need them for annotation purposes. On the other hand, when using all annotations, we have found that nearly 20% of the GOSlims leafs could benefit from enrichment. We believe that this portraits the inevitable lag between knowledge creation and its integration into the ontology. Automated annotation techniques account for over 97% of the total annotations, but due to the general lower confidence researchers have in them, they are frequently disregarded from studies. However, since they greatly increase GO's coverage and their quality is increasing [4], more attention is being directed towards their use. We believe that providing candidate terms to cover areas mainly dedicated to electronic annotations may boost their utility and usage.

5 Conclusions

We have presented an outline for the automated enrichment of the Gene Ontology based on text mining and ontology alignment. We have also identified 17 areas of GO that may benefit from automated enrichment ('hotspots'). These areas have strong electronic annotation activity, but most are not the focus of GO curators. Consequently, we believe that extending these areas would be beneficial, to help GO curators and to support electronic annotation efforts and researchers whose field is not currently one of the areas of interest of GO curators.
Future work will focus on enriching these 'hotspots' using text mining and ontology alignment techniques to support automated enrichment. With this, we hope to provide biomedical researchers with an extended version of GO that can be used 'as is' or by GO developers as a starting point to enrich GO.

6 Acknowledgements

This work was supported by FCT, through the project PTDC/EIA/67722/2006, the Multiannual Funding Programme, and the PhD grant SFRH/BD/42481/2007.

References

1. M. Bada, R. Stevens, C. Goble, Y. Gil, M. Ashburner, J. Blake, J. Cherry, M. Harris, and S. Lewis. A short study on the success of the gene ontology. *Journal of Web Semantics*, 1(1):235–240, 2004.

2. G. Bisson, C. Ndellec, and D. Caamero. Designing clustering methods for ontology building - the mok workbench. In *In Proceedings of the ECAI Ontology Learning Workshop*, pages 13–19, 2000.
3. O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Brief Bioinform*, 7(3):256–274, September 2006.
4. E. Camon, D. Barrell, E. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, and R. Apweiler. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, 6(Suppl 1):S17, 2005.
5. E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32:D262, 2004.
6. A. Faatz and R. Steinmetz. Ontology enrichment with texts from the www. In *In Semantic Web Mining, WS02*, 2002.
7. D. Faure and C. N. Edellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *In LREC workshop on*, 1998.
8. GO-Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue):D258–D261, 2004.
9. M. Hearst. *Automated Discovery of WordNet Relations.* MIT Press, 1998.
10. D. P. Hill, J. A. Blake, J. E. Richardson, and M. Ringwald. Extension and integration of the gene ontology (go): Combining go vocabularies with external vocabularies. *Genome Res.*, 12(12):1982–1991, December 2002.
11. A. Hotho and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence research*, 24:305–339, 2005.
12. J. Jannink and G. Wiederhold. Ontology maintenance with an algebraic methodology: a case study. In *In. Proceedings of AAAI workshop on Ontology Management*, page http://wwwdb. stanfo, 1999.
13. J. B. Lee, J. J. Kim, and J. C. Park. Automatic extension of gene ontology with flexible identification of candidate terms. *Bioinformatics*, 22(6):665–70, Mar 2006.
14. C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, April 2008.
15. C. Roux, D. Proux, F. Rechenmann, and L. Julliard. An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. In *In position paper in Proceedings of the ECAI2000 Workshop on Ontology Learning(OL2000*, 2000.
16. S. Staab. Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In *In Proceedings of the Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*, 2005.
17. T. F. V Parekh, J Gwo. Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In *International Conference of Information and Knowledge Engineering*, 2004.
18. R. G. Valarakos, G. Paliouras, V. Karkaletsis, and G. Vouros. A name-matching algorithm for supporting ontology enrichment. In *In Proceedings of SETN04, 3rd Hellenic Conference on Artificial Intelligence*, pages 381–389. Springer Verlag, 2004.