GOAT: THE GENE ONTOLOGY ANNOTATION TOOL

MICHAEL BADA[1], ROBIN MCENTIRE[2], CHRIS WROE[1], ROBERT STEVENS[1]
{mbada|cwroe|robert.stevens}@cs.man.ac.uk, Robin.A.McEntire@gsk.com

[1]University of Manchester
Department of Computer Science
Kilburn Building, Oxford Road
Manchester M13 9PL UK

[2]GlaxoSmithKline
Upper Merion
709 Swedeland Road
King of Prussia, PA 19406 USA

The Gene Ontology (GO), a structured controlled vocabulary of over 15,000 terms, is becoming the *de facto* standard for describing gene products in terms of their molecular functions, biological processes in which they participate, and the cellular locations in which they are active. However, current annotation editors do not constrain the choice of GO terms users may enter, potentially resulting in inconsistent or even nonsensical descriptions of gene products. Relying upon a DAML+OIL version of GO, including mined GO-term-to-gene-product-type and GO-term-to-GO-term associations, and the FaCT reasoner, GOAT aims to guide the user in the annotation of gene products with GO terms by displaying those field values that are appropriate based on previously entered terms. This will result in annotations of a higher quality, which in turn will facilitate biomedical e-Science.

Introduction

For years, life scientists have been conducting experiments that yield large amounts of complex, dynamic data. The life-science community has become more aware that they lack representations and tools to help marshal the variety of data and higher-level knowledge needed to ask sophisticated questions and perform analyses of these data. In response to this, the Gene Ontology (GO) [1] (Figure 1), a structured controlled vocabulary of over 15,000 terms, has been (and is being) developed to describe the gene products of various organisms, for which it is becoming the *de facto* standard. GO is divided into three subontologies of terms (most of which also have natural-language definitions) which may be used to annotate gene products in terms of their molecular functions, biological processes in which they are involved, and the cellular locations in which they are active. Each term of each of these subontologies is related to its respective parent term(s) via *is-a* or *is-a-part-of* relationships.

Although GO provides a large vocabulary of terms from which to choose to annotate gene products, the three subontologies are (purposely by the GO Consortium) independent of each other, and thus, there are no links between terms of different subontologies. It is possible (though unlikely) that an annotator, in describing a protein, could willfully associate the terms "viral life cycle", "amino-acid biosynthesis" and "extracellular matrix" to that protein; it is more likely that he would accidentally do so. In either case, this is biologically nonsensical. Good annotation relies upon the domain expertise of the annotator and the usability of the annotation tool. We seek to improve upon the latter by creating formal relationships between pairs of GO terms (as well as between GO terms and gene-product types) mined from biological databases and building an application that, relying upon these relationships, will dynamically retrieve and present only those GO terms that are applicable based on the GO terms and the gene-product type already entered by the user.

Such conceptual annotations are not only useful for bioinformaticians querying online data resources and analysis tools but are also a requisite for the kind of activities being undertaken by the UK e-Science programme. Projects such as myGrid [2] are moving towards the Semantic Grid or Information Grid, in which semantic and service rich layers are built on top of the classic Grid where bioinformatics services will reside. The semantic markup of the content and services in that layer is a vital part of ensuring that projects such as myGrid have the appropriate semantic content. Bioinformatics is becoming well-placed to provide such semantically marked up resources, but there is a need to have intelligent tools to facilitate the process of marking up and ensuring the highest possible quality of markup. In this paper we introduce the Gene Ontology Annotation Tool (GOAT) project, which aims to use Description Logic and associated reasoning [3] to guide the annotation process.

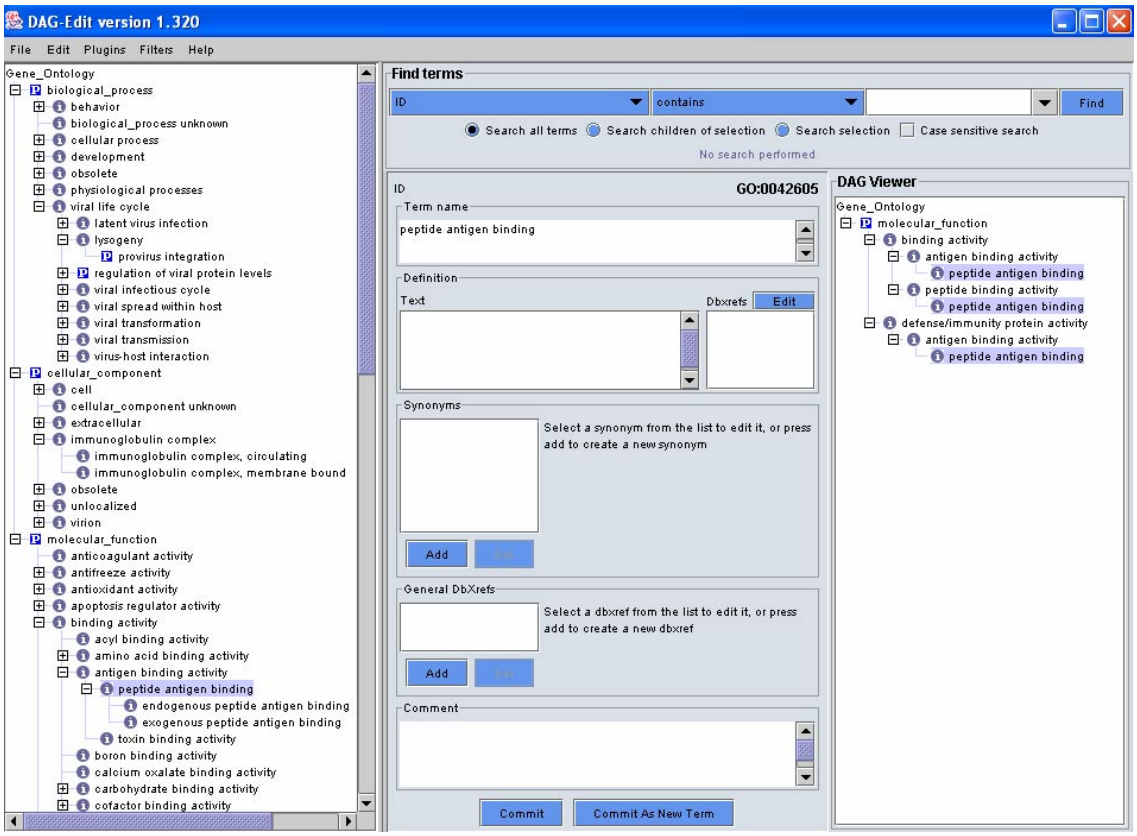

Figure 1. A screenshot of a portion of GO in DAG Edit, a tool for editing controlled vocabularies that are represented as directed acyclic graphs. A term related to its parent via an *is-a* relationship is shown with a circled "i" to the left of the term, and a term related to its parent via an *is-a-part-of* relationship is shown with a squared "p" to the left of the term.

Approach

Currently, the process of annotating gene products with GO terms is an unguided one, in which the user relies upon his domain knowledge to tediously wade through large, often irrelevant parts of GO subtrees in search of the concepts he wishes to use. GOAT seeks to aid the user by presenting him with those terms of the appropriate GO subontology that we have found to be formally associated in GO-associated databases with GO terms he has already entered for the gene product in question. For example, if he has already entered "triplet codon-amino acid adaptor" as the molecular function of a specific tRNA, GOAT could offer "protein biosynthesis" (as well as all of its subconcepts, as the user may want to choose a more specific term) as plausible choices for the GO biological process in which this tRNA participates, since "protein biosynthesis" has been used in combination with "triplet codon-amino acid adaptor" in at least one other credible tRNA database entry. If he then chooses "protein biosynthesis" as the biological process and indicates that he wishes to enter a GO cellular component, GOAT could in turn retrieve all GO cellular-component terms that have been formally associated with "triplet codon-amino acid adaptor" *and* with "protein biosynthesis" in tRNAs. Thus, the suggested terms become more specific (more likely to be accurate) as more information is entered. By offering the most likely terms, the user has a better chance of finding the term(s) he wishes to use.

GOAT is closely related to (and relies heavily upon) another project at the University of Manchester named GONG (Gene Ontology Next Generation) [4]. The goal of GONG is to convert the present GO into a Description-Logic-based ontology and then to further enrich it with formally represented biological knowledge. The former entails translating the ontology into DAML+OIL [5], which is the specific Description Logic we are currently

using. Furthermore, it was necessary to create new concepts under which to classify terms that are related in GO to their parent terms only by *is-a-part-of* relationships since all terms of a formal ontology (apart from the root node(s)) must be connected to one or more parent terms via *is-a* relationships. We have begun to add further semantic content by finding recurring patterns in GO terms (*e.g., x* metabolism, where *x* is some biological chemical) and adding dynamically generated DAML+OIL definitions to these terms.

It is to this DAML+OIL version of GO that we add our GO-term restrictions. The first type of these restrictions is that between GO terms. Currently, the only relationships between terms available from the original GO are the child-parent links that are explicitly represented and the descendant-ancestor links that can be inferred by traversing the hierarchy. Most of the relationships we are adding (in the form of formal DAML+OIL restrictions) are links between the currently independent subontologies of GO. These will, for instance, make connections between terms representing molecular activities and the biological processes with which those molecular activities have been associated. For example, a term representing the molecular activity "hexokinase" can be linked to the biological-process term representing "glycolysis". We are also adding relationships between pairs of GO terms within the same subontology provided that one of the terms does not subsume the other term of the pair (*i.e.*, one is not an ancestor or descendant of the other).

The cross-ontology relationships we are adding are being mined from the complete version of GOA (Gene Ontology Annotation) [6], a database holding all GO-code annotations of Swiss-Prot entries. Specifically, for each GO-term/GO-subontology pair, we determine a set of GO terms each of which is used as a nonelectronically inferred annotation along with the given GO term for at least one Swiss-Prot entry. Terms are added to the set (or replace terms in the set) such that each of the terms is not subsumed by any of the other terms of the set. For the set of a GO-term/GO-subontology pair in which the subontology is the one in which the GO term is placed, each associated term must additionally not subsume or be subsumed by the given term.

The number of restrictions mined from GOA will already result in a significant increase in the size of our DAML+OIL version of GO. Representing only these top-level associated GO terms will minimize the ontology's growth, as GOAT will use a reasoner to retrieve all terms subsumed by the intersection of an entered set of top-level terms. Thus, representing associated terms that are subsumed by other associated terms would be redundant. Examples of GO-term-to-GO-term restrictions can be seen in Figure 2, in which the DAML+OIL GO is shown in OilEd [7], a DAML+OIL-ontology editor. Using only the DAML+OIL descriptions of this figure (and not the many other descriptions of the ontology for the sake of this example), if a user entered "microtubule" as a cellular component and/or "structural molecule" as a molecular function, GOAT would query a reasoner, which would return "microtubule-based movement" as a biological-process term that has been associated with (either of) these concepts. The reasoner would then subsequently be queried to retrieve all descendants of "microtubule-based movement" for display to the user as plausible choices for biological-process annotation.

The second type of these added restrictions is that between GO terms and gene-product types (*i.e.*, types of biological molecules), which were obtained from the various prominent organismal databases that use GO terms to annotate their gene-product entries (*e.g.*, the *Saccharomyces* Genome Database (SGD) [8]). The entries of most of these databases do not have structured fields that classify them into types of biological molecules, and thus, there is no easy way to automatically mine for this type of association. Instead, the databases were manually searched and examined, resulting in a set of zero or more associated gene-product types for each GO term. We assumed that proteins can be annotated with almost any GO term and instead concentrated on finding terms associated with tRNAs, mRNAs, snRNAs, and snoRNAs. These types of macromolecules have more restricted functions (and processes and cellular locations) that can be used to pare a given GO subontology down to a more manageable size for presentation to the user.

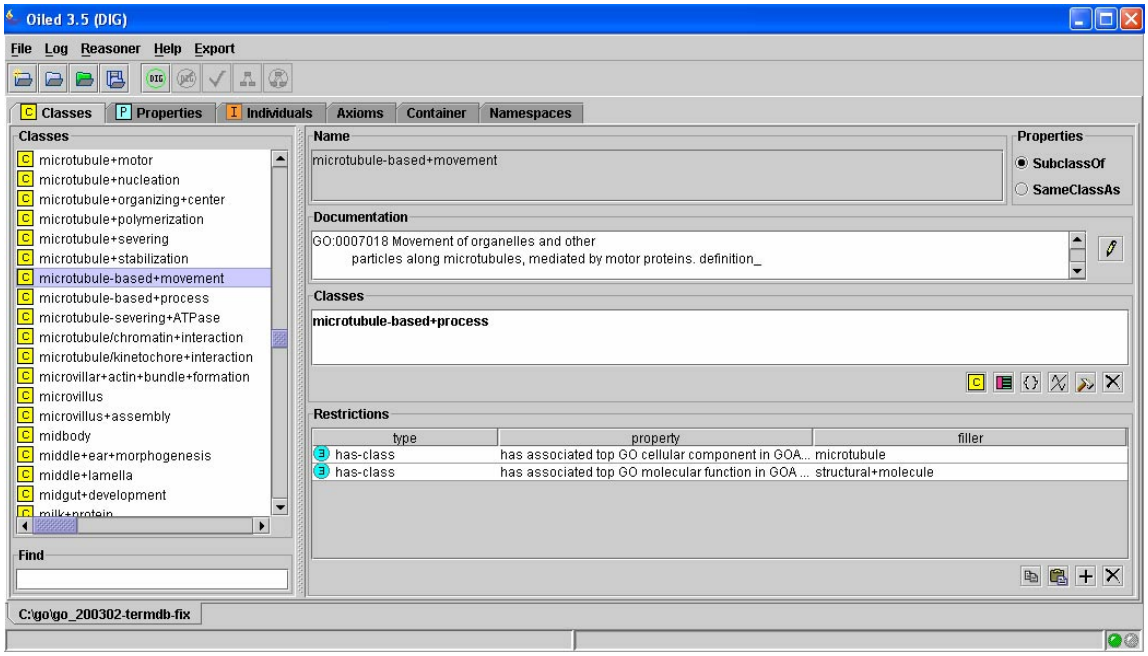

Figure 2. A screenshot of the DAML+OIL version of GO, augmented with restrictions, loaded into OilEd, an editor of DAML+OIL ontologies. The term "microtubule-based movement" has been selected in the left pane, and its two entered restrictions can be seen in the lower right pane, one associating "microtubule" as a cellular component and the other associating "structural molecule" as a molecular function.

After being translated into the DIG file format [9], the ontology, augmented with associations, is loaded into FaCT [10], a classifier of DAML+OIL ontologies, which can then be used to query the ontology. We are currently implementing the user interface of GOAT in PEDRo [11], a simple knowledge-acquisition tool originally designed for use by the proteomics community. PEDRo relies upon an XML Schema as the representation for its GUI, which can be edited to create data-entry forms for any domain. Current work on PEDRo involves extending it with Java classes that can query FaCT for the subsumed terms of dynamically constructed DAML+OIL descriptions. With these extensions, as the user adds terms describing the function, process, and location of a given gene product, these very choices will dynamically restrict the choices offered for the various fields of the form for that gene product. While GOAT is designed specifically for GO-term annotation, PEDRo is a generic tool, and these extensions may be suitably modified to use ontology sources in other domains.

Discussion

Translation to a Description Logic and augmentation with formal term definitions and relationships among the terms will result in a richer, more consistent GO that is open to machine reasoning. Tools driven by this formally represented knowledge can then be built to guide users in specific tasks. Such an example is GOAT, which will use this DAML+OIL version of GO to guide biomedical researchers in the annotation of gene products with GO terms. Specifically, we plan to guide these users by presenting those terms that are most appropriate to enter for a given field given the values that have been entered or chosen for the fields at that point. Currently, life scientists lack such tools and instead must largely rely upon their own expertise and slow traversal of large subhierarchies of GO.

Figure 3 shows a screenshot of GOAT as currently implemented in PEDRo. PEDRo exports entered data as an XML file; each such file consists of a set of gene-product annotations. Each gene-product annotation in turn includes a natural-language name, a gene-product type (*i.e.*, type of biological macromolecule), and a set of one or more GO terms for each of the three GO subontologies. It is these GO-subontology fields where users will be aided in the process of annotation. For example, as shown in Figure 3, the user has entered "triplet codon-amino acid adaptor" in

the molecular-function field and "protein biosynthesis" in the biological-process field and has chosen "transfer RNA (tRNA)" from the menu of gene-product types. Upon indicating that she wishes to enter a value for the cellular-component field, GOAT will dynamically construct a DAML+OIL query asking for those terms of the GO cellular-component subontology that are formally represented in our DAML+OIL GO as being associated with all of this information. Here, this corresponds to the intersection of GO cellular-component terms that have been formally associated with "triplet codon-amino acid adaptor" and "protein biosynthesis" in GOA and with "transfer RNA (tRNA)" in at least one of the examined organismal databases. This subset of GO terms is shown as a tree in a pop-up window, from which she may choose one or more terms as field values. Thus, instead of searching for the appropriate term(s) through all of GO, the user is presented with only the most likely values in the familiar form of GO's hierarchical structure.

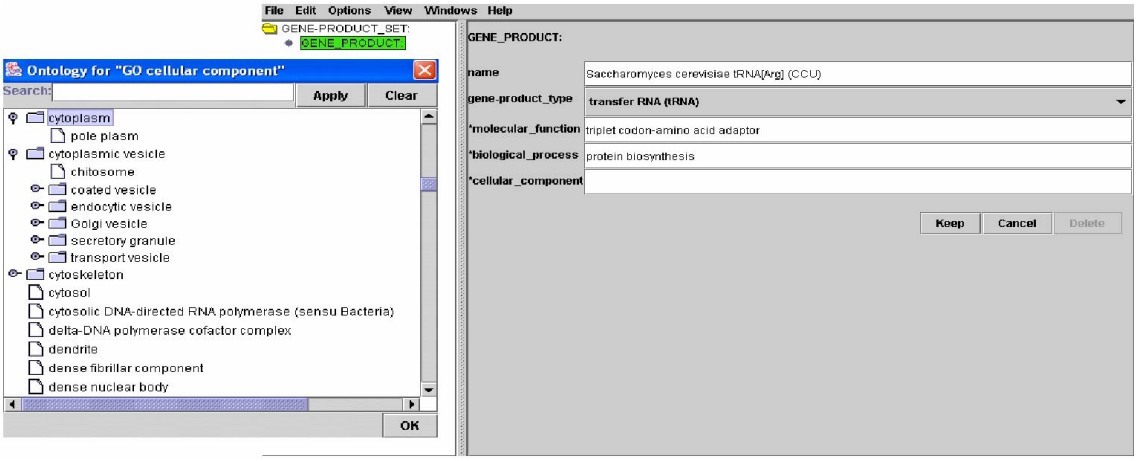

Figure 3. A screenshot of GOAT as implemented in PEDRo. The right window shows that for each gene-product annotation, in addition to a natural-language name field and an enumerated list of gene-product type, there is a field for each of the three subontologies of GO, each of which may have one or more GO terms. Here, the user has indicated that she wishes to enter a value for the field named "cellular_component", and thus, the left window pops up, containing all relevant GO-cellular-component terms based on the GO terms that have already been entered in the other fields.

Other GO browsers, such as QuickGO [12], have also mined the GO-related databases for associations which are then displayed in their interfaces. GOAT goes a step further by the use of reasoning. Simple mined links have to be followed, and it can be difficult to keep track of the many associations that a given term may have. The dynamic use of subsumption reasoning ensures not only that appropriate terms both close and far from a given term are found but that these terms additionally satisfy all constraints implied by information already entered for the gene product. This should make the process of finding relevant terms more efficient and effective.

The semantic annotation of Web or Grid content or services is a vital part of making e-Science function effectively and efficiently. Currently, the annotation process is entirely conducted by humans. In biology, as well as elsewhere, the possibilities for annotation are legion, opening up the possibility of misannotation. We have the technology that can use Description-Logic ontologies and reasoning to help guide the user through the annotation process by "predicting" the next annotation for a gene product based on the set of annotations he has already entered for the gene product. In addition to resulting in more biologically appropriate annotations, this should make the process less tedious and more satisfying by drastically reducing the amount of verbiage through which the user must to wade to reach the appropriate terms. Improving the quality of semantic annotation and easing its production will thus facilitate the progress of e-Science.

Acknowledgments

GOAT is funded by the DTI and EPSRC under the e-Science programme via ESNW. We would also like to acknowledge the collaboration of our industrial partner GlaxoSmithKline.

Bibliography

[1] The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**: 25-29.

[2] Stevens, R., Robinson, A., and Goble, C. A. (2003). myGrid: Personalised Bioinformatics on the Information Grid. *Proceedings of Intelligent Systems in Molecular Biology (ISMB)*, Brisbane, Australia, July 2003.

[3] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., eds. (2003). *The Description Logic Handbook*. Cambridge University Press.

[4] Wroe, C., Stevens, R., Goble, C., and Ashburner, M. (2003). A Methodology to Migrate the Gene Ontology to a Description Logic Environment using DAML+OIL. *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB)*, Hawaii, USA, January 2003.

[5] Horrocks, I.. (2002). DAML+OIL: a Reason-able Web Ontology Language. *Proceedings of Extending Database Technology (EDBT) 2002*, Prague, the Czech Republic, March 2002.

[6] Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., and Apweiler, R. (2003). The Gene Ontology Annotation (GOA) Projection: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research*, **4**, 662-672.

[7] Bechhofer, S., Horrocks, I., Goble, C., and Stevens, R. (2001). OilEd: a Reason-able Ontology Editor for the Semantic Web. *Proceedings of KI2001, Joint German/Austrian Conference on Artificial Intelligence*, Vienna, Austria, September 2001.

[8] Issel-Tarver, L., Christie, K. R., Dolinski, K., Andrada, R., Balakrishnan, R., Ball, C. A., Binkley, G., Dong, S., Dwight, S. S., Fisk, D. G., Harris, M., Schroeder, M., Sethuraman, A., Tse, K., Weng, S., Botstein, D., Cherry, J. M. (2002). *Saccharomyces* Genome Database. *Methods of Enzymology*, **350**, 329-346.

[9] http://dl-web.man.ac.uk/dig/2003/02/dig.xsd

[10] Horrocks, I. (1999). FaCT and iFaCT. In: Lambrix, P., Borgida, A., Lenzerini, M., Möller, R., and Patel-Schneider, P., eds. *Proceedings of the International Workshop on Description Logics (DL'99)*, Linköping, Sweden, July-August 1999.

[11] Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D., Stead, D. A., Yin. Z., Deutsch, E. W., Selway, L., Walker, J., Riba-Garcia, I., Mohammed, S., Deery, M. J., Howard, J. A., Dunkley, T., Aebersold, R., Kell, D. B., Lilley, K. S., Roepstorff, P., Yates, J. R. III, Brass, A., Brown, A. J. P., Cash, P., Gaskell, S. J., Hubbard, S. J., and Oliver, S. G. (2003). A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, **21**(3), 247-254.

[12] http://www.ebi.ac.uk/ego/