The Lexical Properties of the Gene Ontology (GO)

Alexa T. McCray, Allen C. Browne, Olivier Bodenreider

National Library of Medicine
Bethesda, Maryland
{mccray, browne, olivier}@nlm.nih.gov

The Gene Ontology (GO) is a construct developed for the purpose of annotating molecular information about genes and their products. The ontology is a shared resource developed by the GO Consortium, a group of scientists who work on a variety of model organisms. In this paper we investigate the nature of the strings found in the Gene Ontology and evaluate them for their usefulness in natural language processing (NLP). We extend previous work that identified a set of properties that reliably identifies natural language phrases in the Unified Medical Language System (UMLS). The results indicate that a large percentage (79%) of G0 terms are potentially useful for NLP applications. Some 35% of the GO terms were found in a corpus derived from the MEDLINE bibliographic database, and 27% of the terms were found in the current edition of the UMLS.

INTRODUCTION

The Human Genome Project has resulted in a vast amount of data in a relatively short period of time. In addition to the many databases that have been created to store the raw data as soon as they are available, domain ontologies have been developed by a number of groups in order to better manage, compare and interpret the data [1]. The characteristics and coverage of these ontologies vary according to the purposes for which they have been designed.

Blois distinguishes several layers of knowledge in biomedicine, from elementary particles at the most basic level to cellular structures, to the entire organism and, finally, to communities of organisms at the highest levels [2]. Existing terminologies in biomedicine tend to emphasize the middle and higher levels in Blois' hierarchy (e.g., there are many terms for organism pathologies, drugs, devices and procedures to treat these pathologies, etc.), though there are some exceptions. The concepts needed at the molecular level are beginning to be represented by efforts such as the Gene Ontology (GO) initiative.

GO is a construct developed for the purpose of annotating molecular information about genes and their products. [3-5]. The ontology is a shared resource developed by the GO Consortium, a group of scientists who work on a variety of model organisms. The developers are interested in creating a resource that will allow for interoperability among genomic databases and that can be used irrespective of the particular organism being studied. The resource is expected to grow in coverage and to continue to evolve as the research community's understanding of molecular biology increases.

From a nomenclature point of view, the goals of the group are quite pragmatic:

"…the GO Consortium members have chosen to initially focus on three precise sets of terms that are of immediate and exceptional utility to the researcher." [4:1426].

At the same time, the authors have a larger goal in mind: "…the effort described here is an essential start to creating a shared language of biology." [4:1426].

The Unified Medical Language System (UMLS) developed and maintained by the U.S. National Library of Medicine interrelates some sixty terminologies in the biomedical domain. Several of these, including MeSH (Medical Subject Headings) and SNOMED, (Systematized Nomenclature of Medicine) contain some terminology at the basic cellular level, but none is specifically designed for molecular biology. The UMLS Semantic Network has relevant semantic types, including, for example, 'Molecular Function', 'Gene or Genome', 'Nucleotide Sequence', and 'Molecular Biology Research Technique', but additional concepts and semantic types would be needed in order to adequately represent the knowledge in the domain.

Natural language processing (NLP) applications require access to extensive domain knowledge in

order to accurately analyze natural language text [6-8]. The purpose of this study is to analyze the lexical properties of the terms represented in the Gene Ontology in order to determine whether they are suitable for use in NLP applications. Although the ontology has been created for other purposes, it is, nonetheless, a potential source of domain knowledge for NLP applications. The extent to which NLP applications are able to take advantage of the knowledge represented in GO depends in part on the extent to which GO expresses its concepts as well-formed natural language strings.

The work reported here extends our previous work in identifying lexical properties that filter out ill-formed strings in existing biomedical terminologies [9]. Subsequent to our initial work, we created a version of the UMLS MRCON file that, for each string, included a vector of all the properties we had identified. We have made this new file (MRNLP) available as a resource through the UMLS Knowledge Source Server [10].

METHODS

We downloaded the Gene Ontology from the GO web site [11] in February 2002. The file contains 11,381 term records. Of these, 10,366 represent preferred terms, 1,015 represent synonyms of preferred terms, and one is a root term. The terms are grouped into three separate categories: molecular function, biological process, and cellular component. The total number of terms in each category together with some examples are shown below:

Molecular Function (5626 terms)
single-stranded DNA binding
G-protein chemoattractant receptor
palmitoyl-[acyl-carrier protein] hydrolase

Biological Process (4677 terms)
blood vessel development
post-translational membrane targeting
bis (5'-nucleosidyl) oligophosphate metabolism

Cellular Component (1077 terms)
plastid outer membrane
dosage compensation complex
flagellar basal body, MS ring (sensu Bacteria)

Each category represents its own ontology and is organized into isa or part-of hierarchies. An example from the molecular function hierarchy is shown in Figure 1 below:

vasoactive intestinal polypeptide receptor

Gene Ontology
-- molecular function
---- signal transducer
------ receptor
-------- transmembrane receptor
---------- G-protein coupled receptor
---------- secretin-like receptor
---------- vasoactive intestinal polypeptide receptor

Figure 1. GO hierarchy for the term "vasoactive intestinal polypeptide receptor", showing that it is a type of transmembrane receptor

We analyzed the GO terms using a variety of methods. These included attempting to map GO terms to the UMLS, applying our NLP filter to the full set of GO terms, searching for the terms in a corpus of MEDLINE abstracts, and investigating whether the terms and their constituent words appeared in the SPECIALIST lexicon.

Our first step was to map the GO terms to the UMLS. We used the 2002 edition of the UMLS [12] and Version 2.0 of the UMLS Knowledge Source Server [10]. The mapping was constrained by the UMLS semantic types and the categories of the source GO terms. For example, a mapping from a function term in GO was only considered successful if it mapped to a UMLS concept having a compatible semantic type.

Our earlier work on the lexical properties of natural language strings resulted in the development of a possible NLP filter [9]. We used a statistical model to determine which lexical properties would be most useful for identifying strings that are likely to appear in biomedical text. The resulting filter was based on a collection of the most promising lexical properties and was intended primarily for use in filtering the UMLS Metathesaurus for natural language processing purposes.

In this study, we applied the NLP filter to GO, and also individually to all the UMLS constituent vocabularies because we were interested in knowing not only how well the GO terminology would fare, but also how it compared to other biomedical terminologies. The filter is shown below.

(((NB_SOURCES > 1) or (NB_WORDS <= 5)) and (not CT_NON_ALPHANUM))

A string passes the filter if it appears in more than one source or if it consists of five or fewer words,

and if it consists only of alphanumeric characters. (Exceptions are made for hyphens, which are treated as spaces in the comparison, and for apostrophes, which often signal the possessive, though this latter leads to a small amount of noise in the results.)

The filter expresses a number of insights. The longer a string is, the less likely it is to be found as a unit in natural language text. Further, if a string appears in several sources created by different groups, then it is more likely to reflect a standard way of expressing a concept. Finally, if the string contains characters such as parentheses, commas, or other non-alphanumeric characters, the less likely it is to be an item found in well-formed text.

Because in some cases there are multiple versions of the same vocabulary (e.g. COSTAR 1989, 1992, 1993 and 1995 releases are all included in the UMLS), we collapsed these multiple versions into a single "family" of sources and used the family in computing the source count. Since GO is not yet integrated in the UMLS, if a string mapped into the Metathesaurus, it automatically passed the first part of the filter (number of sources > 1).

Next, we searched for the GO terms in a corpus derived from MEDLINE. The corpus consists of titles and abstracts for over 400,000 citations entered into MEDLINE during 1999. We matched each string against the corpus, retaining all string features, (e.g., punctuation, spacing, word order) with the exception of case. The appearance of a string in a corpus is one indicator of its usage as a natural language term, although the converse is not necessarily true. That is, there are various reasons why a perfectly good bit of language may not appear in a corpus. The corpus may not be large enough, or its scope may be too narrow.

Finally, we searched for the GO terminology in the SPECIALIST lexicon. The lexicon contains syntactic information, such as part of speech and inflectional patterns for each of the lexical items it contains. The lexicon is a curated resource that contains well-formed general English and biomedical terminology. We were interested in how many of the GO terms were exact matches, as well as how many of the individual words in multi-word terms were found .

RESULTS

The primary results of our investigation are summarized in Table 1 below.

	Molecular Function	Biological Process	Cellular Comp.	Total Terms
GO strings	5626	4677	1077	11380
In UMLS	2436	256	370	3062 (27%)
Passed NLP filter	4338	3730	907	8975 (79%)
In corpus	2125	1318	570	4013 (35%)
Full term in lexicon	636	166	204	1006 (9%)

Table 1 . Profile of Gene Ontology Strings

The table gives results for the terms in each of the three GO categories. The percentages in the table reflect the percentage of the total GO terms that have the property. For example, 4013 terms were found in the corpus. This represents 35% (4013/11380) of the total GO terms.

27% of the GO terms matched into the 2002 version of the UMLS. The majority of the terms that matched are found in either MeSH or SNOMED or both. The number of matching terms by category in the two vocabularies is shown in Table 2 below.

	Molecular Function	Biological Process	Cellular Comp.
MeSH	2269	164	331
SNOMED	1119	86	161

Table 2. GO terms matching MeSH and SNOMED

Some 43% (2436/5626) of the molecular function terms were found in the UMLS, while only 5% (256/4677) of the biological processes were found. This may be explained in part by the high degree of specificity of the biological process terms, as well as by the fact that many of the processes refer to specific organisms (e.g., while the cell component "mitotic spindle" is found in the UMLS, neither of the GO biological process terms "mitotic spindle assembly" nor "mitotic spindle assembly (sensu Saccharomyces)" is found).

79% of the total GO terms passed the NLP filter. This result places GO in the top third of the UMLS vocabularies with regard to the well-formedness of its strings. Figure 2 below shows the distribution of the UMLS constituent vocabularies.

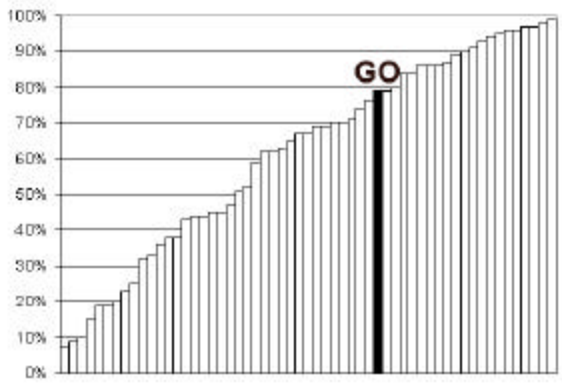

Figure 2. Distribution of UMLS vocabularies with respect to the percentage of the terms that passed the NLP filter.

The vocabularies are arranged with the leftmost vocabulary having the lowest percentage of terms passing the NLP filter and the rightmost vocabulary having the highest percentage. Other vocabularies that scored as highly or higher than GO include the World Health Organization Adverse Drug Reaction Terminology, the Vanderbilt University Canonical Clinical Problem Statement System, and DSM3R (Diagnostic and Statistical Manual of Mental Disorders). Vocabularies that scored somewhat lower than GO include SNOMED, MeSH, and ICD-9-CM (International Statistical Classification of Diseases, Clinical Modification).

Some examples of GO terms that passed the NLP filter, as well as examples of those that did not pass, are shown below.

Example GO terms

Passed the NLP filter:
nuclear pore membrane protein
Epstein Barr Virus-induced receptor
DNA strand elongation

Did not pass the NLP filter:
G2/M transition of mitotic cell cycle
1-phosphatidylinositol-4-phosphate kinase, class IA
peptide:N-glycanase

Only 4,013 GO terms were found in the MEDLINE 1999 corpus. This is surprisingly low when compared with the count of the terms that passed the NLP filter. It is not immediately clear why this would be the case. The relatively small size of the corpus may be a contributing factor, as well as the highly specific nature of some of the terminology. Some examples of GO terms not found in the corpus are listed below.

Sample terms not found in the MEDLINE corpus
high affinity zinc uptake transporter
cell wall mannoprotein biosynthesis
transcription factor TFIIIB

A small percentage (9%) of full GO terms was found in the lexicon. In some cases these are single word terms, such as "aminopeptidase", "nucleus", and "lysosome", and in other cases, they are multi-word phrases, such as "DNA repair enzyme", "blood coagulation factor", and "reverse transcriptase".

A much larger percentage of the words that make up GO terms was, however, found in the lexicon. A total of 4,780 unique words comprise the 11,380 GO terms. Of these, 3,558 (74%) are found in the SPECIALIST lexicon. The largest number of the words found are nouns, e.g., "cell", "regulation", "growth"; a smaller number are adjectives, e.g.., "nuclear", "inorganic", "mitochondrial"; some are words that can be used as both nouns and verbs, e.g. "transport", "repair", "cluster", "control", "release", "damage"; and a much smaller number represent other parts of speech such as prepositions and conjunctions.

2,664 of the GO words appear at least twice in GO terms (with a large number appearing in hundreds of terms). Of these 2,664 words, 83% (2,219) are also in the SPECIALIST lexicon. The twenty most frequent words (excluding numbers and prepositions) found in GO terms are shown below in descending order of frequency:

Twenty most frequent GO words

protein (741)	kinase (254)
receptor (630)	peptidyl (244)
metabolism (591)	cell (228)
biosynthesis (518)	complex (216)
catabolism (424)	factor (197)
transporter (412)	DNA(183)
acid (394)	sensu (181)
transport (376)	amino (171)
binding (290)	synthase (163)
dehydrogenase (255)	phosphate (161)

The word frequency list presents another view of the nature of the specialized domain covered by the GO terminology.

CONCLUSIONS

Several conclusions can be drawn based on the results presented in this study. First, the Gene Ontology is suitable as a resource for natural language processing applications. The ontology has been carefully developed and contains a large number of interrelated concepts in the domain.

The percentage of terms that passed our NLP filter compares favorably with a number of vocabularies represented in the UMLS, and a significant number of the words in GO terms are represented in the SPECIALIST lexicon, making this lexical resource, together with all of its tools readily accessible to NLP researchers [13].

Less than one third of the GO terms matched existing strings in the UMLS. There may be at least two explanations for this. It may be that, even though a particular string does not match, the concept actually exists in the UMLS in some other form. It is also likely, however, that the concept does not yet exist in any of the constituent UMLS vocabularies because it represents a newly emerging concept in the field.

Research in biology, and, in particular, in cellular and molecular biology, is advancing more rapidly than it was possible to predict even a decade ago. This fact coupled with the enormous advances in information technology during the same time period have created a situation in which masses of new data are created daily. This has led research groups, such as the Gene Ontology Consortium, to collaborate in the development of tools and processes to manage the flood of data being generated. Scientists who may previously have worked independently in their own research communities have come to see the advantage of working together in pursuit of a common goal. The development of the Gene Ontology has arisen in such an environment.

In this study we were interested in determining the nature of this cooperatively developed ontology, in particular with regard to its lexical properties. The methodology and results presented here extend our earlier work in developing methods to assess biomedical terminologies for natural language processing purposes.

REFERENCES

1. Stevens R, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Briefings in Bioinformatics 2000 Nov; 1(4):398-414.
2. Blois MS. Information and medicine: The nature of medical descriptions. Berkeley: University of California Press., 1984:38-63.
3. Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 2000 May; 25:25-9.
4. Gene Ontology Consortium. Creating the Gene Ontology resource: design and implementation. Genome Research 2001; 11:1425-33.
5. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G, Sethuraman A, Weng S, Botstein D, Cherry JM. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Research 2002;:30(1):69-72.
6. Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA Annu Fall Symp 2001;:181-5.
7. Hahn U, Romacker M, Schulz S. How knowledge drives understanding – matching medical ontologies with the needs of medical language processing. Artif Intell Med. 1999;15(1):25-51.
8. Rzhetsky A, Koike T, Kalachikov S, Gomez SM, Krauthammer M, Kaplan SH, Kra P, Russo JJ, Friedman C. A knowledge model for analysis and simulation of regulatory networks. Bioinformatics 2000; 16(12):1120-8.
9. McCray AT, Bodenreider O, Malley JD, Browne AC. Evaluating UMLS strings for natural language processing. Proc AMIA Annu Fall Symp 2001;:448-52.
10. UMLS Knowledge Source Server. http://umlsks.nlm.nih.gov/. Accessed March 5, 2002.
11. Gene Ontology Consortium Home Page. http://www.geneontology.org/. Accessed March 5, 2002.
12. National Library of Medicine. UMLS Knowledge Sources. 13th edition, 2002.
13. McCray AT. The nature of lexical knowledge. Meth Inf Med 1998;37:353-60.