GO-WORDS: An Entropic Approach to Semantic Decomposition of Gene Ontology Terms

*Tuanjie Tong, Yugyung Lee, and Deendayal Dinakarpandian**

School of Computing and Engineering, University of Missouri-Kansas City, Kansas City, Missouri 64110, USA

ABSTRACT
The Gene Ontology (GO) has a large and growing number of terms that constitute its vocabulary. An entropy-based approach is presented to automate the characterization of the compositional semantics of GO terms. The motivation is to extend the machine-readability of GO and to offer insights for the continued maintenance and growth of GO. A prototype implementation illustrates the benefits of the approach.

1 INTRODUCTION

The underlying motivation of the work described in this paper is to map annotations based on the Gene Ontology (GO) (Ashburner, et al., 2000) to a semantic representation that exposes the internal semantics of GO terms to computer programs. The Gene Ontology (GO) views each gene product as being a structural component of a biological entity, being involved in a biological process, and as having a molecular function. These three dimensions of component (C), process (P) and function (F) are hierarchically refined into several thousand subconcepts or GO terms for a fine-grained description of gene products, and ultimately a representation of collective biological knowledge. The machine-readability of GO is based on explicit IS-A or PART-OF relations between different GO terms (Fig. 1). The representation of each GO term in terms of a phrase in English is primarily meant for human readability, and not machine-readability (Wroe, et al., 2003) (Fig. 1). For example, while both humans and computer programs can understand that 'Folic Acid Transporter Activity' is one kind of 'Vitamin Transporter Activity," only a human reader can appreciate that proteins annotated with 'Folic Acid Transporter Activity' actually *transport* the vitamin *folic acid*. In other words, the compositional semantics embedded within each GO term is not currently accessible by computer programs; each term *per se* is effectively a black box or meaningless string of characters to computer programs.

It has been estimated that about two-thirds of GO terms (Ogren, et al., 2004) contain another GO term as a substring within it. For example, the GO term 'Transporter Activity' is a substring of several GO terms such as 'Vitamin Transporter Activity' and 'Biotin Transporter Activity.' In other words, many GO terms are combinations of distinct semantic units, as opposed to being a completely new concept. The compositional nature of GO terms has the side effect of resulting in a combinatorial increase in the size of GO. For example, 'Folic Acid' appears in 12 different GO terms like 'Folic Acid Transport,' 'Folic Acid Binding,' and 'Folic Acid Transporter Activity.' Similarly, the vitamin Biotin appears in 23 GO terms, including 6 terms identical to that for Folic Acid except for the replacement of 'Folic Acid' with 'Biotin,' e.g., 'Biotin Transport,' 'Biotin Binding' and 'Biotin Transporter Activity.' This phenomenon has been one of the motivating factors behind the GO Annotation Tool (GOAT) (Bada, et al., 2004) and the Gene Ontology Next Generation (GONG) project (Wroe, et al., 2003), which suggested having multiple intersecting hierarchies, with a proposed evolution towards a DAML+OIL representation. Reasons for studying the compositional nature of GO are to suggest missing relations (Mungall, 2004; Ogren, et al., 2004), suggest new terms (Lee, et al., 2006; Ogren, et al., 2004), increase computability of GO (Doms, et al., 2005; Ogren, et al., 2004; Wroe, et al., 2003), and for providing models for GO-based analysis of natural language processing of text (Blaschke, et al., 2005; Couto, et al., 2005; Doms and Schroeder, 2005).

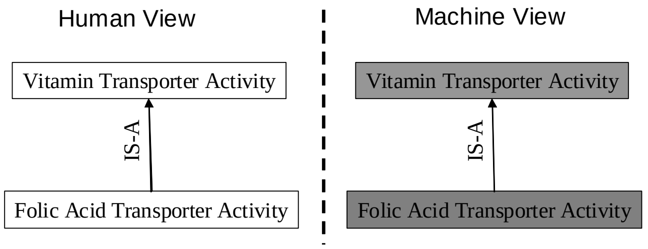

Human View
Machine View
Vitamin Transporter Activity
IS-A
Folic Acid Transporter Activity
Vitamin Transporter Activity
IS-A
Folic Acid Transporter Activity

Fig. 1. The internal semantics of GO terms are visible to humans but not to computer programs

One way to discretize GO is to represent it as a language consisting of progressive concatenation of tokens in the form of regular expressions. An example of this is *Obol* (Mungall, 2004), a language that exploits the regularity of GO term names to represent it in Backus-Naur format. However, this is applicable to only a subset of all GO terms.

In this paper, we use an entropic approach for the analysis of regularity of GO term nomenclature. We show how this

[*] To whom correspondence should be addressed.

may be used to detect sets of GO terms sharing similar semantics. The decomposition of GO terms presented here also suggests a way to minimize the complexity of GO.

2 METHODS

The general principle is to find clusters of GO terms sharing similar semantic structure. Entropy (see below) is used to find GO terms that share consistent location of a specific token (word) within them. Each cluster is evaluated and a corresponding semantic rule created.

Analysis of position-dependent conservation of GO tokens

Each GO term (version Feb 16th, 2006), including synonyms, was tokenized on white space into a sequence of individual words. For example, the GO term “L-amino acid transport” is tokenized as “L-amino” + “acid” + “transport.” Entropy (Shannon 1950) is used to measure the regularity in location of each token within all GO terms:

$$EP_t = \sum_{i=1}^{l} - p_i^t \log p_i^t$$

where EP_t is the positional entropy of token *t*, *l* is the length (in number of tokens) of the longest GO term or synonym that token *t* is observed to occur in, and p_i^t is the probability of finding token *t* at position *i*. If the logarithm is in base 2, then entropy can be quantified in terms of bits. Recognizing that gene product and molecule names embedded in GO terms consist of a variable number of tokens, we choose to note the position of each token relative to *both* the beginning and end of each GO term. For example, the token “acceptor” almost always occurs at the end of a GO term (with the sole exception of the term “electron acceptor activity”). Thus, it is uniformly the first term when counted from the end of a GO term, with a resulting low positional entropy of 0.08 with respect to the end (EPE). In contrast, this token has a highly variable position when counted from the start of a GO term (as many as 15 different locations) resulting in a high positional entropy (EPS) value of 3.3. If we focus only on an EPS value, we would miss its positional conservation, i.e., tendency to occur at the very end of GO terms.

Since Shannon entropy is based only on proportions, it does not distinguish between token distributions like [1, 1] (token found once at the first position, and once at the second) and [100, 100] (token found a hundred times each at the first and second positions). Both would yield an entropy value of 1 bit even though there are only 2 occurrences of the former and 200 of the latter. To distinguish between such tokens, the absolute numbers of occurrence at a given distance from either the start or end of GO terms are also recorded. The calculated entropies are then ‘normalized’ (NEP) by adding 0.1 to the calculated value and dividing by the total number of occurrences. Division of the entropic value by the total number of occurrences yields lower values for a higher number of token occurrences. The addition of 0.1 bit helps to distinguish between tokens having an entropy of zero but differing in their frequency of occurrence within GO terms. For the above examples, this would yield values of (0.1/2 = 0.05) and (0.1/200 = 0.0005) respectively, thus yielding a lower NEP value (implying higher degree of positional conservation after correction for more frequent occurrence) for the more frequent token.

Semantic mapping rule generation

Tokens with low positional entropy, high number of occurrences or low normalized positional entropy are used as a starting point for the generation of rules. For each such token, the corresponding set of GO terms is verified for semantic uniformity and a corresponding rule generated. This takes minimal time as the majority of terms in a set follow the same pattern. For example, ‘binding’ is a token that has much lower entropy when measured from the end (0.28 bit) than from the beginning (2.16 bits). The vast majority, 1544 out of 1597, of GO terms containing the token ‘binding’ end with it. 1524 of these are of the general form ‘Entity’ + ‘binding’ where ‘Entity’ represents one or more tokens in succession representing a single concept. The Entity most often specifies a molecule, and sometimes a structural component. The 20 exceptions include terms like ‘Protein domain specific binding’ and ‘regulation of binding.’ Thus, the discretizing rule applicable to gene products {Mi} annotated with these GO terms may be stated as ‘Mi binds Entity.’ In other words, each corresponding GO term (e.g. Zinc Binding) is decomposed into a relational term (e.g. Binds) and the embedded concept (e.g. Zinc). Thus, if the protein “40S ribosomal protein S27” is annotated with the GO term ‘Zinc Binding,’ then the corresponding discretized semantic form is ‘40S ribosomal protein S27 Binds Zinc.’ Fig. 2 summarizes the general procedure with another example. Triplets of this form correspond to MachineProse assertions (Dinakarpandian, et al., 2006) and can contribute to an incremental knowledge-base distinct from paper publications.

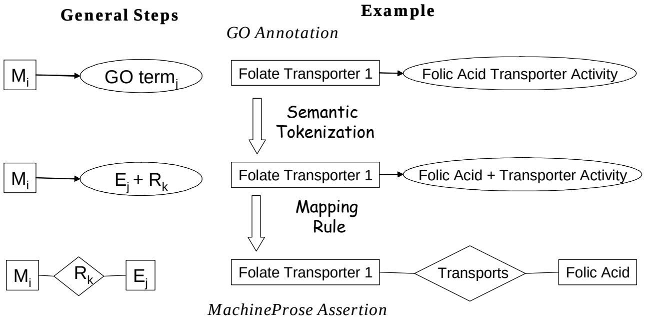
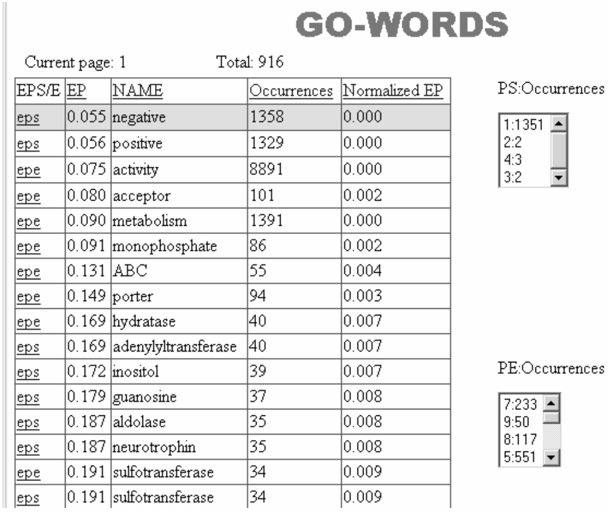


Fig. 2. Mapping a GO annotation to a discretized triplet. The general procedure is shown on the left together with a specific example on the right

3 RESULTS & DISCUSSION

GO-WORDS browser



GO-WORDS

Current page: 1 Total: 916

EPS/E	EP	NAME	Occurrences	Normalized EP
eps	0.055	negative	1358	0.000
eps	0.056	positive	1329	0.000
epe	0.075	activity	8891	0.000
epe	0.080	acceptor	101	0.002
epe	0.090	metabolism	1391	0.000
epe	0.091	monophosphate	86	0.002
epe	0.131	ABC	55	0.004
epe	0.149	porter	94	0.003
epe	0.169	hydratase	40	0.007
eps	0.169	adenylyltransferase	40	0.007
eps	0.172	inositol	39	0.007
eps	0.179	guanosine	37	0.008
eps	0.187	aldolase	35	0.008
eps	0.187	neurotrophin	35	0.008
epe	0.191	sulfotransferase	34	0.009
eps	0.191	sulfotransferase	34	0.009

PS:Occurrences

1:1351
2:2
4:3
3:2

PE:Occurrences

7:233
9:50
8:117
5:551

Fig.3. Browser for analyzing tokens/words found within GO terms. Columns 2 and 5 are measures of positional variation of each token within GO terms, column 1 indicates whether position in each row is with respect to the beginning or end of corresponding GO terms, column 3 shows name of token, and column 4 shows number of GO terms it is found in.

Tokenizing GO resulted in a 9152 unique tokens from a total of 37,403 terms (20115 canonical + 17288 synonym terms). Each token occurred 13.7 times on average. The most frequent token was found to be "activity," occurring a total of 8891 times. In contrast, almost half the tokens (4204), e.g. "xylem," occurred only once. We implemented a browser (Fig. 3) to analyze *position-wise* frequencies and entropy of GO-terms. EP stands for entropy. The suffix S, as in EPS indicates that positions were counted from the beginning of the string, whereas the suffix E, as in EPE, indicates that positions were counted backwards from the end of the string. The prefix N indicates normalization (see Methods above). Each token was analyzed using multiple metrics. For example, Table I shows that the token 'negative' has the lowest positional entropy because it occurs most of the time at the beginning of a GO term (1351 out 1358 occurrences with a corresponding EPS=0.055, and normalized EPS=0.00004). In contrast, the token 'oxidoreductase' (not shown) has the highest positional entropy (EPE=3.854;NEPE=0.019) because its 212 occurrences are spread over 29 different positions within GO terms like '*oxidoreductase* activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced pteridine as one donor, and incorporation of one atom of oxygen.' Clearly, it is potentially easier to map GO terms containing the token 'negative' than 'oxidoreductase' to a machine-readable representation.

The GO-WORDS browser is a useful tool to gain insights into the composition of GO terms. With respect to this paper, we focused on using it mainly to create semantic mapping rules. Thus, tokens with low values of NEPE (observed range=0.00004 – 0.562) (Table I) and a large number of occurrences were used to select GO terms for semantic mapping to an assertion representation.

Given a token and position either from the beginning or end of a string, the GO-WORDS browser lists all GO terms and synonyms that share the token at a given position. For example, the token 'transporter' occurs second from the end (517 out of 650) in GO terms like the following:

name: L-ornithine transporter activity
name: S-adenosylmethionine transporter activity
exact_synonym: S-adenosyl methionine transporter activity
name: adenine nucleotide transporter activity
name: spermine transporter activity
name: sulfite transporter activity

Table I. Tokens with lowest normalized positional entropy

Token	Normalized Entropy	Token	Normalized Entropy
activity	nepe=0.000	dehydrogenase	neps=0.001
negative	neps=0.000	cell	nepe=0.001
positive	neps=0.000	complex	nepe=0.001
metabolism	nepe=0.000	metabolism	neps=0.001
activity	neps=0.000	receptor	neps=0.001
binding	nepe=0.000	biosynthesis	neps=0.001
regulation	neps=0.000	transporter	nepe=0.001
of	neps=0.000	binding	neps=0.001
of	nepe=0.000	formation	neps=0.002
biosynthesis	nepe=0.000	ligand	nepe=0.002
pathway	nepe=0.000	catabolism	neps=0.002
regulation	nepe=0.001	transport	nepe=0.002
formation	nepe=0.001	cell	neps=0.002
anabolism	nepe=0.001	synthesis	neps=0.002
synthesis	nepe=0.001	acid	nepe=0.002
differentiation	nepe=0.001	proliferation	nepe=0.002
catabolism	nepe=0.001	acceptor	nepe=0.002
receptor	nepe=0.001	degradation	neps=0.002
breakdown	nepe=0.001	exocytosis	nepe=0.002
degradation	nepe=0.001	anabolism	neps=0.002

The general pattern for the above examples is "Entity transporter activity." Thus, the mapping rule applicable to gene products {Mi} annotated with these GO terms may be stated as 'Mi transports Entity,' where entity is presumed to be the prefix of 'transporter activity.' This assumption is true in 420 of the 440 cases. Exceptions to the rule include terms like "siderophore-iron (ferrioxamine) uptake transporter activity" and "transporter activity." In the former, only a subset of the prefix of "transporter activity" represents an

Entity, i.e, the word 'uptake' doesn't conform to the same pattern. The latter is the parent term representing the abstract concept of 'transporter activity.'

The GO token entropic measure helps in clustering terms that share a token at the same relative position. Based on the general patterns 'Entity binding' and 'Entity transporter activity,' 23780 and 903 annotations respectively were mapped to discretized triplets. However, the entropic analysis is based on the naïve assumption that each token represents a concept. In reality, names of entities often consist of a variable number of words strung together, e.g., lipoprotein lipase. Measuring the positional entropy of a token from either end helps mitigate this problem to an extent, but only to an extent. In particular, GO terms where the token of interest is flanked by entities of variable length will not show a peak in the positional distribution. Further, since it is based purely on a textual approach (no prior semantics), manual verification is required to find sub-concepts that are made up of contiguous tokens.

4 CONCLUSION

This paper has presented and addressed the advantages of a discretized triplet representation of GO annotations and a partially automated approach for doing so. In future, we intend to extend the approach to the entire Gene Ontology, combine information from other sources, and devise a sophisticated search interface that shall incorporate the MachineProse relation ontology (Dinakarpandian, et al., 2006).

The number of terms in GO has been rapidly growing since its inception (Ashburner, et al., 2000). The total number of terms has grown from 4507 in 2000 to more than 20,000 in Feb 2006 (Gene Ontology Consortium). One reason is a richer description, but redundancy of nomenclature is also a factor. As GO is continuously revised (terms becoming obsolete, renamed and rearranged), maintaining its semantic integrity is quite challenging. This paper suggests an approach to a leaner GO that is both people and machine friendlier by allowing annotations to be built from reuse of semantically defined building blocks. This would lessen the growth rate of GO, with the resultant smaller size helping in ensuring uniformity and semantic consistency of GO. The benefits would be easier maintenance of GO and higher semantic transparency. In the interim, a triplet view of GO annotations offers a pragmatic solution. A potential advantage is to facilitate searches specified as a set of triplets, occupying the middle ground between a natural language interface and a keyword-based one.

Since a large number of entities within GO are general or specific names of molecules, extracting the embedded molecular ontology would be a useful adjunct. Using other ontologies like ChEBI (ChEBI) and completed mappings between GO and other ontologies (Johnson, et al., 2006) would be useful in this regard.

ACKNOWLEDGEMENTS

We would like to acknowledge funding support for this work from the University of Missouri-Kansas City Research Board to DD (FRG 2006) and from the University of Missouri Research Board to YL (UMRB 2005).

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Bada, M., Turi, D., McEntire, R. and Stevens, R. (2004). Using Reasoning to Guide Annotation with Gene Ontology Terms in GOAT. SIGMOD Record, **33**(2).

Blaschke, C., Leon, E.A., Krallinger, M. and Valencia, A. (2005) Evaluation of BioCreAtIvE assessment of task 2, *BMC bioinformatics*, **6 Suppl 1**, S16

ChEBI Chemical Entities of Biological Interest. http://www.ebi.ac.uk/chebi/

Couto, F.M., Silva, M.J. and Coutinho, P.M. (2005) Finding genomic ontology terms in text using evidence content, *BMC bioinformatics*, **6 Suppl 1**, S21.

Dinakarpandian, D., Lee, Y., Vishwanath, K. and Lingambhotla, R. (2006) MachineProse: an Ontological Framework for Scientific Assertions, *J Am Med Inform Assoc*, **13**, 220-232.

Doms, A., Fuche, T., Burger, A. and Schroeder, M. (2005) How to Query the GeneOntology. *Proceedings of Symposium on Knowledge Representation in Bioinformatics*. Finland.

Doms, A. and Schroeder, M. (2005) GoPubMed: exploring PubMed with the Gene Ontology *Nucleic Acids Res.*, W783-W786.

Gene Ontology Consortium. Gene Ontology Annotations. http://www.geneontology.org/

Johnson, H.L., Cohen, K.B., Baumgartner, J.W.A., Lu, Z., Bada, M., Kester, T., Kim, H. and Hunter, L. (2006) Evaluation of Lexical Methods for Detecting Relationships Between Concepts from Multiple Ontologies. *Pacific Symposium on Biocomputing*. 28-39.

Lee, J.B., Kim, J.J. and Park, J.C. (2006) Automatic extension of Gene Ontology with flexible identification of candidate terms, *Bioinformatics*, **22**, 665-67

Mungall, C.J. (2004) Obol: integrating language and meaning in bio-ontologies, *Comparative and Functional Genomics*, 509-520.

Ogren, P.V., Cohen, K.B., Acquaah-Mensah, G.K., Eberlein, J. and Hunter, L. (2004) The compositional structure of Gene Ontology terms, *Pacific Symposium on Biocomputing*, 214-225.

Shannon, C.E. (1950) Prediction and entropy of printed English. The Bell System Technical Journal, **30**, 50-64

Wroe, C.J., Stevens, R., Goble, C.A. and Ashburner, M. (2003) A methodology to migrate the gene ontology to a description logic environment using DAML+OIL, *Pacific Symposium on Biocomputing*, 624-635