The Compositional Structure of Gene Ontology Terms

P.V. Ogren, K.B. Cohen, G.K. Acquaah-Mensah, J. Eberlein, and L. Hunter

Pacific Symposium on Biocomputing 9:214-225(2004)

THE COMPOSITIONAL STRUCTURE OF GENE ONTOLOGY TERMS

P.V. OGREN[a]
University of Colorado at Boulder, Dept. of Computer Science, Boulder, CO;
Center for Computational Pharmacology, University of Colorado Health Sciences Center, School of Medicine, Denver, CO

K.B. COHEN[a], G.K. ACQUAAH-MENSAH,
J. EBERLEIN, L. HUNTER
Center for Computational Pharmacology, University of Colorado Health Sciences Center, School of Medicine, Denver, CO

An analysis of the term names in the Gene Ontology reveals the prevalence of substring relations between terms: 65.3% of all GO terms contain another GO term as a proper substring. This substring relation often coincides with a derivational relationship between the terms. For example, the term *regulation of cell proliferation* (GO:0042127) is derived from the term *cell proliferation* (GO:0008283) by addition of the phrase *regulation of*. Further, we note that particular substrings which are not themselves GO terms (e.g. *regulation of* in the preceding example) recur frequently and in consistent subtrees of the ontology, and that these frequently occurring substrings often indicate interesting semantic relationships between the related terms. We describe the extent of these phenomena—substring relations between terms, and the recurrence of derivational phrases such as *regulation of*—and propose that these phenomena can be exploited in various ways to make the information in GO more computationally accessible, to construct a conceptually richer representation of the data encoded in the ontology, and to assist in the analysis of natural language texts.

1 Introduction

1.1 Motivation

The Gene Ontology (GO) is the result of an effort to enumerate and model concepts used to describe genes and gene products[1,2]. We refer to the central unit of description in GO as a *concept.* Concepts consist of a unique identifier and one or more strings that provide a controlled vocabulary for unambiguous and consistent naming. In this paper, we refer to these strings as *terms.* (Our use of the word *term* subsumes GO names and synonyms, and in that sense is consistent with the use of *term* in the terminology literature (e.g. Jacquemin[3]), although not with GO's use of it.) Concepts exist within a hierarchy of isA and partOf relations in a directed acyclic graph (DAG) that locates all concepts in the knowledge model with respect to their relationships to other concepts. As Wroe et al.[4] and Yeh et al.[5] point out,

[a] These authors contributed equally to the work reported in this paper.

the terms themselves contain additional information that is implicit in the term names, but is not explicitly represented by the isA and partOf relations that constitute the "model" of the ontology. For example, in the term *positive regulation of cell migration* (GO:030335), the facts that (a) the concept encodes a regulation relationship between two entities, and (b) that the direction of the regulation is positive are implicit, but in order to exploit those facts computationally, the term itself must be subjected to linguistic analysis. We are interested in using these sorts of facts to make the information in GO more computationally accessible and to leverage the information implicit in the ontology into a conceptually richer knowledge base. As a step in that direction, we undertook an analysis of the structure of the linguistic content of GO terms. Some hypotheses about the nature of this structure immediately presented themselves:

- Many GO terms seem to contain other GO terms as proper substrings. For example, the term *positive regulation of cell migration* (GO:0030335) contains the GO term *regulation of cell migration* (GO:0030334) as a proper substring.
- In this process of deriving GO terms from other terms, certain strings which are not themselves GO terms seem to recur frequently. For example, the string *regulation of* occurs 1,053 times in GO terms, 330 of these times directly modifying some GO term (as in the above example) to produce a new GO term. We hypothesized that these strings, which in general we refer to as *complements,* might themselves have interesting and exploitable characteristics. For example, it might be the case that all tokens of a particular complement or all terms that are modified by a particular complement might occur within a particular subtree of the GO hierarchy.

In this paper we characterize the extent of these phenomena in GO, both with respect to the inclusion of GO terms in other terms, and with respect to the patterns of usage of the strings that are added to GO terms to produce other terms. We will demonstrate that some of these complements encode specific semantic relations, both corresponding to and more granular than the sanctioned GO relations of isA and partOf. We refer to the subset of complements that constitute these semantically contentful complements as *derivational phrases*. We then give some examples of ways in which insight into these phenomena can be useful.

1.2 What it means to have (compositional) structure

An ontology can include terms which have a semantic relationship (consisting of the relationships between nodes, encoded in the edges that link them within the DAG) but no surface linguistic relationship. For example, the MeSH ontology (`http://www.nlm.nih.gov/mesh/meshhome.html`) contains the following terms that are all related via the semantic relationship isA, but that have no "linguistic" similarity (in terms of the strings that label them):

- Gram-Negative Bacteria [B03.440]
  - Mollicutes [B03.440.560]
    - Mycoplasmatales [B03.440.560.580]
      - Acholeplasmataceae [B03.440.560.580.100]

Conversely, an ontology can also include terms that have semantic relationships that coincide with very clear surface linguistic relationships. For example, GO contains the following set of terms that are all related via the semantic relation partOf. They also have an evident, patterned linguistic similarity, in that each lower node in the hierarchy contains its parent term as a proper substring:

- membrane [GO:0016020]
  - inner membrane [GO:0019866]
    - mitochondrial inner membrane [GO:0005743]
      - mitochondrial inner membrane peptidase complex [GO:0042720]

A central claim of this paper is that terms such as the preceding GO examples possess linguistic relationships with other terms that correspond to the semantic relationships encoded in GO. These linguistic relationships also can be used to uncover other underlying semantic relationships that enrich the GO knowledge model. These relationships are evident in the patterns of inclusion of terms in other terms, and also in the strings that are added to the included terms to form the "including" terms. It will be seen that these linguistic relationships are quite common and that a large majority of them do, in fact, correspond both to the sanctioned GO relationships and to other semantic relations as well.

In this paper we refer to complements that encode systematic semantic relations as *derivational phrases.* This implies that GO curators engage in term construction through a mental process that explicitly represents the "derivational phrases" that we discuss. The Consortium encourages them to do so, directing curators to "Aim to be reasonably descriptive, even at the risk of…verbal redundancy." Whether or not such a mental process exists is a moot point—the facts about terms and complements hold regardless.

2 Methods and results

2.1 Incidence of inclusion of terms in other terms

As our corpus we used the XML-formatted version of the June 2003 release of the Gene Ontology[b]. This version of GO contains 13,361 concepts, associated with 13,361 names and 3090 synonyms for those names, for a total of 16,451 terms.

[b] `go_200306-termdb.xml`, downloadable from `www.godatabase.org`.

We examined all terms for the occurrence of other terms within them, including their own synonyms. We counted all occurrences of any term within another term. For every such occurrence, we classified the nature of the relationship between the two terms in a number of ways. We categorized the type of edge (i.e., isA, partOf, or synonymy) between the two concepts within the ontology's DAG. We counted instances where there was and was not a dominance relation between the two nodes, and where there was a dominance relation, we counted instances where it was entirely along isA edges, entirely along partOf edges, and where it was along a combination of the two. Also, where a dominance relation existed, we counted separately instances of dominance and instances of immediate dominance[c]. We also classified the directionality of the substring relation between the two strings—when a dominance relation existed, we counted separately instances where the superior node's term was included within the inferior node's term, and where the inferior node's term was included within the superior node's term. (Intuitively, you would not expect to find cases of the latter.) We also counted instances of a term being included in one of its own synonyms. In total, we found that 65.3% (10,747/16,451) of all GO terms contain another GO term. These terms correspond to 72.2% (9,658/13,361) of all GO concepts. Table 1[d] shows the distribution of the terms that contain other terms between the two relations (isA and partOf) and the two possible directions of containment (child or inferior term contains parent or superior term vs. parent contains child). (The rows do not sum up because a term can appear in multiple rows. For example, *jasmonic acid mediated signaling pathway (induced systemic resistance)* (GO:0009864) contains the term *jasmonic acid mediated signaling pathway* (GO:0009867) to which it is related by immediate dominance via the isA relation, so it appears once on the *isA, $A < B$, $A \subset B$* row (and also on the *isA, $A << B$, $A \subset B$* row). It also contains *induced systemic resistance* (GO:0009682), to which it is related as a partOf by immediate dominance, and so it also appears on the *partOf, $A < B$, $A \subset B$* row (as well as the row for the corresponding transitive relation)).

Of these terms that contain other terms, most are related transitively by isA and are inferior nodes that contain their superior nodes—52.5% of all GO terms fit this

---

[c] Following Partee et al., we define *dominance* and *immediate dominance* as follows: "We say that a node *x dominates* a node *y* if there is a connected sequence of branches in the tree extending from *x* to *y*. This is the case when all the branches in the sequence have the same orientation away from *x* and toward *y*.…If *x* and *y* are distinct, *x* dominates *y*, and there is no distinct node between *x* and *y*, then *x immediately dominates y*"[6]. We use the mathematical terminology of dominance, rather than the parent/child/ancestor/descendant usage of the Consortium, because we found it to allow a clearer and more compact exposition in the *Methods* section.

[d] Additional related data for this and subsequent tables is available at `http://compbio.uchsc.edu/Hunter_lab/Ogren/psb2004.html`.

description, with 25.5% of all GO terms containing the node that is immediately superior to them via the isA relation. We found some instances of containment in intuitively unlikely directions. 25 terms (0.15%) contain their immediate isA descendant, e.g. *mating* (GO:0007618) isA *mating behavior* (GO:0007617), and *memory* (GO:0007613) isA *learning and/or memory* (GO:0007611). 14 terms (0.07%) are wholes that have as a substring one of their parts, e.g., *ribosome biogenesis* (GO:0007046) partOf *ribosome biogenesis and assembly* (GO:0042254).

Table 1 Occurrence of terms within other terms The "combined" rows are for cases where the dominance relation involves both isA and partOf edges. "Sibling" rows are for terms that are both immediately dominated by the same node. A < B indicates A immediately dominates B, A << B indicates A dominates B, A ⊂ B indicates A is a proper substring of B.

	percentage of all GO terms
isA, A < B, A ⊂ B	25.5% (4,197/16,451)
isA, A << B, A ⊂ B	52.5% (8,639/16,451)
isA, A < B, B ⊂ A	0.15% (24/16,451)
isA, A << B, B ⊂ A	0.15% (24/16,451)
partOf, A < B, A ⊂ B	3.65% (601/16,451)
partOf, A << B, A ⊂ B	4.1% (673/16,451)
partOf, A < B, B ⊂ A	0.07% (12/16,451)
partOf, A << B, B ⊂ A	0.07% (12/16,451)
combined, A << B, A ⊂ B	8.01% (1318/16,451)
combined, A << B, B ⊂ A	0%
sibling/isA	0.84% (139/16,451)
sibling/ partOf	0%
synonym	0.84% (139/16,451)
no dominance relation	16.8% (2,763/16,451)
total instances of terms containing another term	65.3% (10,747/16,451)

In every case where one term contained the other as a proper substring, we recorded the phrase that was the complement of the substring with respect to the superstring, classifying these derivational phrases in a variety of ways as well which we describe below.

2.2 Characteristics of complements

Whenever one term contained another as a proper substring, we recorded the complement of the substring with respect to the superstring. Note that we do not claim that every instance of a substring relationship between two terms correlates with a non-trivial semantic relation between them—a key part of the analysis must be to attempt to find a principled way to differentiate between trivial and non-trivial ones. Two characteristics of the complements collected were examined—the frequency of occurrence and the consistency of their usage. The collection of complements includes 9,799 types and 16,915 tokens.[e] We found that 7,686 (78.4%) of the types occurred only once. While these complements may correspond to important semantic relationship between the pairs of terms involved, we did not include them in the data presented in this section. We examined two subsets of the complements for their consistency of usage; those that occurred twice or more and those that occurred five times or more. The former contains 2113 (21.6%) of the complement types comprising 9229 (54.6%) of the complement tokens, while the latter contains 361 (3.7%) of the complement types comprising 4705 (27.8%) of the complement tokens.

Consistency of complement usage was examined in two ways. First, the dominance relations encoded in GO between the pairs of terms associated with each complement type were noted. Second, the locations in the ontology of where each complement occurs were collected and summarized. Table 2 summarizes the resulting data on the consistency of complements to particular relation types. To generate the data in this table, we listed all of the complements associated with a particular dominance relation, and then counted the number of them that only occurred within that relation. For example, the data in *isA, $A < B$, $A \subset B$* row indicates that there are 462 complement types found in pairs of terms related via immediate dominance in the isA hierarchy with frequency greater than or equal to two, and that 293 of those types only occur in pairs of terms related via immediate dominance by an isA edge. In general, complements do tend to be consistent with a particular type of relation. Those complements that *are* consistent with a particular relation may have the status of derivational phrases—that is, they add consistent semantic content to the terms to which they are appended to produce new terms. Some implications of this finding are discussed in section 3 below.

Table 2 Specificity of complements to relation types

	Freq >= 2	Freq >=5
isA, $A < B$, $A \subset B$	293/462 (63.4%)	41/88 (46.6%)
isA, $A << B$, $A \subset B$	1470/1655 (88.8%)	187/282 (66.3%)
partOf, $A < B$, $A \subset B$	27/38 (71.1%)	3/5 (60.0%)

[e] A *type* is a unique complement. A *token* is an instance of occurrence of that complement. *Positive* and *negative* are two types; *positive* occurs as a complement 380 times, so there are 380 tokens of the type *positive.*

partOf, A << B, A ⊂ B	37/49 (75.5%)	3/5 (60.0%)

Table 3 summarizes the resulting data on the consistency of complements with respect to particular locations in the GO hierarchy. Formally, we count all complements for which it is the case that there is some node *n* in the DAG such that all tokens of that complement are dominated by *n*, and all complements for which it is the case that there is some node *n* such that all of the terms that are modified by that complement are dominated by *n*. When either all tokens of a complement occur under a common node, or all of the terms that are modified by a particular complement occur under a common node, this may be a strong indicator that the complement encodes a semantically significant relation. The indication is stronger the lower in the hierarchy that the common node is, so we also differentiate between cases where the common node is one of the three root nodes (biological process, molecular function, and cellular component), and cases where the common node is lower than one of the three roots. Again, the cells do not sum up since a particular complement can occur in multiple rows. Overall, the data show that the metric of occurrence under a common node is very effective at narrowing the list of likely derivational phrases from among the total set of complements—only 13.8% (1,354/9,799) of complements have more than one token and also share an ancestor—while still proposing a usefully large set of potential derivational phrases.

Table 3 Occurrence of complements under a common node Each line can be read 'Total of complement types ….' Only complements with frequency greater than or equal to two were considered.

with frequency > = 2	2,113
with a shared ancestor	1,354
in isA under common node	1,240
in isA under common node below roots	1,182
in isA whose terms are under common node	1,178
in isA whose terms are under common node below roots	996
in partOf under common node	52
in partOf under common node below roots	52
in partOf whose terms are under common node	52
in partOf whose terms are under common node below roots	52

3 Implications and conclusions

The data that we present above are consistent with the hypothesis that derivation of GO terms from other GO terms is a widespread phenomenon, and that this phenomenon often involves particular phrases that are used repeatedly to indicate particular semantic relations. We see a number of applications for this insight into the structure of GO terms. These applications include assistance in the evaluation

and curation of the GO ontology; converting the information encoded in terms into a computable form; and applying GO to problems in natural language processing.

3.1 Aids to the evaluation and curation of GO

The high-frequency occurrence of certain complements suggests that these complements might themselves be suitable GO concepts or relationship types. For example, the string *regulation of* is one of the most common complements, occurring by itself as a complement in 330 terms (and occurring again in 313 terms as a substring of *positive regulation* and in *314* terms as a substring of *negative regulation*). However, there is no GO concept for regulation, per se. The frequent use of this term suggests that perhaps it should be. This suggestion is supported by other work on GO terms, which is consistent with the idea that word frequency is a good indicator of suitability for inclusion in the ontology. McCray et al. give a list of the twenty most common words found in GO terms (the top ones being *protein, receptor, metabolism, biosynthesis,* and *catabolism,* along with fifteen other words that are clearly related to the domain of molecular biology)[7]. We found that if you allow for the addition of the word *activity* to words like *receptor,* then 80% of their top ten words and 55% of their top twenty words are themselves GO terms. Frequent usage within GO terms seems to correlate reasonably well with suitability for termhood.

The derivational phrases also point us towards potential GO concepts when they occur unexpectedly. The "expected" pattern for the derivational phrases that we found is that they directly modify a GO term (to produce a new GO term). Occasionally, they occur in conjunction with a string that is not itself a GO term, and in such situations, that string itself possibly should be a GO term. For example, the derivational phrase *negative regulation of* is often followed immediately by a GO term. However, we noticed a GO term, *negative regulation of REM sleep* (GO:0042322), which is notable in that the string that follows *negative regulation of*, i.e. *REM sleep*, is not itself a GO term. This alerts us to the possibility that *REM sleep* should be added to the set of concepts in GO, probably as a child of *sleep*, GO:0030431. We collected all strings that occurred in such contexts—i.e., they are modified by a derivational phrase but are not themselves GO terms—and counted their occurrences. Two of the authors with biological expertise independently reviewed a list of all such strings that occurred six or more times and rated each string as "would be a good novel GO term" or "would not be a good novel GO term." In 22.2% of the cases (24/108) they concurred that the string would be a good novel GO term. The importance of this for a knowledge engineering effort is that the domain experts only had to consider just over 100 terms to discover 24 new terms for the ontology. Examples of strings that they concurred on include *dehydrogenase activity, methylation,* and *stroma.* We recommend these as strong

candidates for inclusion in the ontology.[f] In a number of additional cases, they indicated that the strings should probably be synonyms for existing terms. For example, the string *envelope* occurred on this list. It appeared following ten different complements that in other cases are followed immediately by an embedded GO term, such as *nuclear, viral, inner,* etc. It is not itself a GO term, but the biologists suggested that it should probably be added as a synonym to *external encapsulating structure* (GO:0030312). Thus, investigating the derivational structure of GO terms has helped us uncover new concepts that are good candidates for inclusion in the ontology and to find appropriate synonyms for concepts that are already in the ontology.

Derivational phrases can also point us towards cases where two concepts that are already in the ontology should be related, but aren't. For example, we found that the string *limonene* occurs as a complement, and when it occurs (as a complement per se), it usually occurs in isA relations in the biological process ontology. The one exception to this is its occurrence in *limonene monooxygenase activity* (GO:0019113), which has a proper substring *monooxygenase activity* (GO:0004497). This seems like an omission, and we suggest that such an edge should be added between these two concepts.

3.2 Enriching GO's conceptual representations

We are interested in leveraging GO into a conceptually richer, more interconnected knowledge base[8]. The relations directly encoded in the isA and partOf relations of GO are an excellent starting point. In the context of the Gene Ontology Next Generation project, Wroe et al. suggest starting by adding the relations *part_of_cellular_component, part_of_molecular_function,* and *part_of_biological_process.* They also relate GO entries to external data sources, and elaborate the representation of metabolic processes[4]. Yeh et al. suggest relations that encode organism or taxon specificity (which they parse out of the terms themselves), macromolecular structure, and temporality[5]. Williams and Anderson suggest relations that encode temporality and location, among others[9].

We would like to also be able to exploit the information that is in the terms themselves; as Wroe et al. have pointed out, "Biologists are able to interpret information…within term names….However, this implicit information is inaccessible to computer applications"[4]. The derivational phrases themselves suggest many relations. One such relation is seen when the derivational phrase encodes the fact that the contained term is the object of the process named by the containing term. For example, *nucleosome* (GO:0000786) is a substring of *nucleosome disassembly* (GO:0006337). There is no link between them in GO.

[f] The 24 candidates were submitted to the GO consortium.

The semantic relationship between them is that the nucleosome undergoes the process of disassembly. We can represent this by means of an *undergoesProcess* edge in a semantic network, or by a similarly labelled slot in a frame-based representation or predicate in a description logic; the fact that the relation exists at all is suggested to us by awareness that the string *disassembly* occurs as a complement. More elaborate examples of the suggestion of relations or slot values by the occurrence of derivational phrases occur, as well. For example, the strings *positive*, *regulation of*, and *positive regulation of* are all very frequent derivational phrases. Together, they suggest two regulation-oriented relations. One is *regulatedItem*, and the other is *regulationDirection*. For instance, *positive regulation of mitotic cell cycle* (GO:0045931) would be represented in a conceptually richer knowledge base with a *regulatedItem* slot whose value was *mitotic cell cycle* (GO:0000278) and a *regulationDirection* slot whose value was *positive*.

Table 4 Relation names suggested by frequently occurring complements Each example complement type has the frequency with which it occurred as well as a token '*term*' indicating where the contained term is located in the containing term. For example, the term *negative gravitaxis* (GO:0048060), fits into the complement pattern 'negative *term* 388' where 388 is the count of all the terms that fit this pattern.

Relation Name	Example Complement Types
Regulation direction	Negative *term* 388, positive *term* 380
Process type	*term* binding 30, *term* biosynthesis 35
Base type	Purine *term* 55, pyrimidine *term* 53
Oxygen availability	Aerobic *term* 18, Anaerobic *term* 26
Substance type	Protein *term* 26, peptide *term* 12
Chirality	d-*term* 24, l-*term* 18
Activity type	*term* binding activity 23
Cellular location	Nuclear *term* 23, mitochondrial *term* 22
Gender	Female *term* 15, male *term* 14
Amino acid type	Serine *term* 14, glycine *term* 14
Substance affinity level	High affinity *term* 14, low affinity *term* 12
Cell division	*term* mitotic 12, *meiotic* term 11
Development stage	Adult *term* 11, larval *term* 11

We examined some of the most frequently occurring complements to determine possible relations that they suggest. Table 4 contains a partial list of relations based on this analysis that could be incorporated into GO. Each row has one or more example complement types used as evidence for the usefulness of the proposed

relation. Tanabe found some of these, e.g. *regulation direction* and *cellular location*, to be relevant to text data mining in the molecular biology domain[10]. While these suggested relations could be named and modeled in numerous ways, they indicate productive avenues for future ontological development.

3.3 Natural language processing

The observation that GO terms can contain other GO terms also points us towards a solution to the problem of recognition of variant forms of terms in natural language texts. Non-statistical approaches to this problem, such as the National Library of Medicine's MetaMap[11] and Jacquemin's FASTR system[3], tend to perform well, but at the computational expense of performing extensive linguistic analysis of both the terminology itself and the natural language text. We suggest that noting the inclusion patterns of terms within other terms allows us to find meaningful linguistic boundaries purely on the basis of comparisons between terms within the ontology, without the necessity of submitting the terms to further morphological or syntactic analysis. An example of this approach can be seen with respect to the problem of dealing with coordination, or linkage of phrases by words like *and, or,* and *but not.* Consider the following sentence from Gilmore and Romer[12]: *These findings suggest that FAK functions in the regulation of cell migration and cell proliferation.* If one were indexing this sentence by GO terms, the best matches would be *regulation of cell migration* (GO:0030334) and *regulation of cell proliferation* (GO:0042127). The first one is easy—the challenge is to get the second one, without being misled into matching instead to *cell proliferation* (GO:0008283). Though space does not allow a full description of our approach, we have been successful in handling this and similar examples by licensing the recognition of discontinuous embedded terms and their associated complements when the discontinuity is due to the intervention of a conjunction and another GO term.

3.3 Conclusion

We have shown that substring relations between terms are prevalent in GO, and that complementary phrases in the superstrings recur frequently. Specificity of complements to relation types and occurrence of complements under common nodes allow us to differentiate between derivationally meaningful substring relations and incidental ones, as well as their associated complement phrases. Awareness of these phenomena can be used to make the information encoded in GO terms more computationally accessible, to assist in the curation of GO, to leverage GO into a conceptually richer knowledge base, and to analyze natural language texts.

Acknowledgements

The authors gratefully acknowledge support for this research from a grant from the Wyeth Genetics Institute and from NIH/NIAAA grant U01-AA13524-02. This work would not have been possible without the tremendous investment of time and effort by the GO Consortium into the development of the Gene Ontology. We also thank the anonymous reviewers whose comments led to an improved paper.

References

1. Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology" *Nature Genetics* 25:25-29 (2000)
2. Gene Ontology Consortium, "Creating the Gene Ontology resource: design and implementation" *Genome Research* 11:1425-1433 (2001)
3. Jacquemin, Christian, *Spotting and discovering terms through natural language processing* (The MIT Press, USA, 2001)
4. Wroe, C.J.; Stevens, R.; Goble, C.A.; and M. Ashburner, "A methodology to migrate the Gene Ontology to a description logic environment using DAML+OIL" *Pacific Symposium on Biocomputing 2003*
5. Yeh, Iwei; Karp, Peter D.; Noy, Natalya F.; and Russ B. Altman, "Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO)" *Bioinformatics* 19(2):241-248, 2003.
6. Partee, Barbara H. ; ter Meulen, Alice; and Robert E. Wall, *Mathematical methods in linguistics*, corrected first edition (Kluwer Academic Publishers, 1993)
7. McCray, Alexa T.; Browne, Allen C.; and Olivier Bodenreider, "The lexical properties of the Gene Ontology (GO)" (*Proceedings of the AMIA 2002 Annual Symposium*, pp. 504-508)
8. Acquaah-Mensah, George K.; Eberlein, Jens; McGoldrick, Daniel J.; Fox, Lynne M.; Cohen, K. Bretonnel; and Lawrence Hunter, *An evaluation metric for molecular biology knowledge-bases* (UCHSC Center for Computational Pharmacology Technical Report TR-03-01, 2003)
9. Williams, Jennifer and William Anderson "Bringing ontology to the Gene Ontology" *Comparative and Functional Genomics* 4:90-93, 2003.
10. Tanabe, Lorraine, *Text mining the biomedical literature for genetic knowledge* (George Mason University doctoral dissertation, 2003)
11. Aronson, Alan R., "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program" (*Proceedings of the AMIA Symposium 2001*, pp. 17-21)
12. Gilmore, A.P. and L. H. Romer, "Inhibition of focal adhesion kinase (FAK) signaling in focal adhesions decreases cell motility and proliferation." *Molecular Biology of the Cell* 7(8):1209-24.