*Implications of Compositionality in the Gene Ontology for Its Curation and Usage*

P.V. Ogren, K.B. Cohen, and L. Hunter

Pacific Symposium on Biocomputing 10:174-185(2005)

# IMPLICATIONS OF COMPOSITIONALITY IN THE GENE ONTOLOGY FOR ITS CURATION AND USAGE

PHILIP V. OGREN
*University of Colorado at Boulder, Dept. of Computer Science, Boulder, CO USA*

K. BRETONNEL COHEN, LAWRENCE HUNTER
*Center for Computational Pharmacology, University of Colorado Health Sciences Center, School of Medicine, Aurora, CO USA*

In this paper we argue that a richer underlying representational model for the Gene Ontology that captures the implicit compositional structure of GO terms could have a positive impact on two activities crucial to the success of GO: ontology curation and database annotation. We show that many of the new terms added to GO in a one-year span appear to be compositional variations of other terms. We found that 90.2% of the 3,652 new terms added between July 2003 and July 2004 exhibited characteristics of compositionality. We also examine annotations available from the GO Consortium website that are either manually curated or automatically generated. We found that 74.5% and 63.2% of GO terms are seldom, if ever, used in manual and automatic annotations, respectively. We show that there are features that tend to distinguish terms that are used from those that are not. In order to characterize the effect of compositionality on the combinatorial properties of GO, we employ finite state automata that represent sets of GO terms. This representational tool demonstrates how ontologies can grow very fast, and also shows that small conceptual changes can directly result in a large number of changes to the terminology. We argue that the curation and annotation findings we report are influenced by the combinatorial properties that present themselves in an ontology that does not have a model that properly captures the compositional structure of its terms.

## 1. Introduction

There have been several papers in recent years that address the need to redesign the underlying model for representing concepts in the Gene Ontology [1,2]. For example, Verspoor et al. propose mapping GO to a lexical semantic network [3]. The Gene Ontology Next Generation project [4] seeks to represent GO using description logics, while Yeh et al. [5] show how GO could be maintained in the frame-based Protégé environment. Joslyn et al. [6] suggest representing GO as a poset. The Open Bio-Ontology Language (OBOL) [7] describes GO terms by means of a Prolog grammar that helps generate and maintain logical definitions. These

efforts at suggesting varying representational schemes for GO have been productive. For example, a number of these efforts have detected missing terms and relations. The Gene Ontology Annotation Tool [8] makes use of description logic constructs to help facilitate annotation of gene and protein databases by use of constraints that help human annotators choose multiple concepts from different GO sub-ontologies that are logically consistent. Verspoor et al. [9] and Joslyn et al. [6] demonstrate some of the benefits that can accrue in natural language processing and high-throughput gene expression data analysis from other representations. In this paper we do not promote or propose any specific model but instead present experimental data that suggests that any new representation of GO should explicitly model the compositional nature of GO terms, which in the current incarnation of the ontology is only implicit.

### 1.1. *Compositional Structure of GO Terms*

GO terms exhibit underlying compositional structures that are not represented explicitly in the ontology. In previous work, we demonstrated that GO terms often contain other GO terms as proper substrings [10]. For example, *activated T-cell proliferation* (GO:0050798) contains the term *T-cell proliferation* (GO:0042098), and these embeddings can continue through several levels of ontological structure. In our earlier work, strings like *activated* that are added to other terms to form a new term were labelled *complements;* the term that is found inside another term is labelled a *contained term*. Where our previous work focused on the complements that are added to previously existing terms to create new terms, work by Verspoor et al. [3] looked closely at the terms to which complements are added, and showed that insights into the relations between these terms can be gained by considering the various derivational processes by which new terms are created. Taken together, these two papers address compositionality from two complementary perspectives and establish that compositionality is a pervasive and prevalent phenomenon in GO.

## 2. Methods

### 2.1. *Assessing implications of compositionality for curation*

To test the hypothesis that the compositional nature of GO terms significantly contributes to the growth of GO, we examined the 3,652 new terms added between July 2003 and July 2004. Our methods for detecting compositional terms are based

on the observation that composed terms will be very similar to other terms, or will have contained terms, or both. Since no single measure will detect every compositional term, we performed two separate experiments to estimate the number of new terms added to GO in the past twelve months that appear to be compositional. For both experiments we compared the new terms with previously existing terms as well as other new terms.

The first experiment is based on the observation that compositionally derived terms are very similar to other terms (see Section 2.3). We measured similarity by calculating the minimum edit distance (MED) between each new term and all other terms in the ontology (including other new terms). The MED is the minimum number of edits needed to convert one string into another. An edit can be either an insertion, a deletion, or replacement of one string by another [11]. While typically this algorithm is applied on the character level to compare two words, we applied the algorithm on the word level to compare terms. We used the Levenshtein weighting [12], in which insertion, deletion, and substitution all have equal weights of one. With this weighting, the terms *trophectoderm cell proliferation* (GO:0001834) and *natural killer cell proliferation* (GO:0001787) have a MED of two because a transformation from the former to the latter can be accomplished with one string replacement (*trophectoderm* → *natural*) and one insertion (∅ → *killer*). We then counted the number of new terms that are at a MED of exactly one from some other term[a]. Of the 3,652 new terms added to GO between July 2003 and July 2004, we found that 2,696 or 73.2% of the new terms had a MED of exactly one from at least one other term.

The second experiment looks for terms that contain another term as a proper substring (a string $x$ is a proper substring of the string $y$ if all of $x$ is in $y$ and $x$ is not identical to $y$) [13]. For example, *ligase activity, forming nitrogen-metal bonds, forming coordination complexes* (GO:0051002) contains the term *ligase activity, forming nitrogen-metal bonds* (GO:0051003). The longer term, which is a new term, is clearly compositionally related to the shorter term, but would not have been detected by the MED test, since the MED between them is three. We found that 2,507 or 68.6% of the new terms contained another term.

---

[a] With this weighting, all pairs of two-word terms, e.g. *symporter activity* (GO:0015293) and *hydrolase activity* (GO:0016787) have a minimum edit distance of one. However, in these cases the low minimum edit distance probably does not reflect a derivational relationship, so we calculated the minimum edit distance only when one of the terms is longer than two words.

We then determined the intersection and the union of the sets of terms discovered by the two experiments. Table 1 summarizes the results of the two experiments and the union and intersection of the two sets of terms. The union of the sets of terms discovered by both experiments contains 3,294 terms or 90.2% of the new terms. Thus, a large majority of the new terms have at least one characteristic of compositional terms. The number of terms identified by both measures is 1,909 or 52.5% of the new terms. Thus, a small majority of the new terms have two characteristics.

| | Count | Percentage |
|---|---|---|
| Total New Concepts | 3652 | 100% |
| MED1 | 2696 | 73.2% |
| CT | 2507 | 68.6% |
| MED1 _ CT | 1909 | 52.5% |
| MED1 ∪ CT | 3294 | 90.2% |
| Neither | 358 | 9.8% |

Table 1. MED1 – new concepts that have a minimum edit distance of one from another term. CT – new concepts that have a contained term.

### 1.2. *Assessing implications of compositionality for annotation*

#### Annotations per term

Analysis of how GO is used to annotate gene and protein databases reveals that much of GO is either barely used or not used at all. We downloaded all annotations available at the GO website and counted for each GO term how many annotations (or usages) are associated with it[b]. We split the annotations into three broad categories: human curated, computer generated, and all annotations combined[c]. For each GO term we counted the number of annotations that contain that term. We then ranked the GO terms based on their frequency of usage. Figure 1 shows the data for the

[b] The annotations were downloaded on July 5, 2004 from http://www.geneontology.org/GO.current.annotations.shtml.

[c] For the human curated annotations we used evidence codes IC, IDA, IEP, IGI, IMP, IPI, ISS, and TAS. For automatically generated annotations we used the evidence code IEA. For the *all* category we included every evidence code including NAS, NR, and ND.

manually curated annotations on both absolute and logarithmic scales[d]. Table 2 highlights some of the data across all three categories.

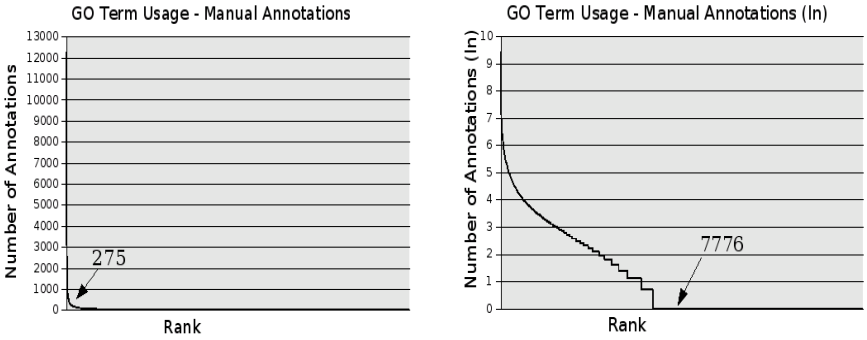


Figure 1. Distribution of manually curated annotations. These graphs are histograms that show for each rank how many annotations there were. The graph on the left gives the data on an absolute scale. The 275th most frequently used term is used in 200 annotations and is pointed out in the graph. The graph on the right gives the data on a logarithmic scale. Terms ranked 7,776 or higher are not used at all.

| Usage Count (ln) | Usage Count | Manual n | % | Automated n | % | All n | % |
|---|---|---|---|---|---|---|---|
| 0 | 0 or 1 | 9,568 | 57.3% | 9,272 | 55.5% | 8,044 | 48.2% |
| (0-2) | 2-7 | 2,876 | 17.2% | 1,286 | 7.7% | 1,602 | 9.59% |
| [2-4) | 8-54 | 3,378 | 20.2% | 2,590 | 15.5% | 3,069 | 18.4% |
| [4-6) | 55-401 | 768 | 4.60% | 2,259 | 13.5% | 2,603 | 15.6% |
| [6-∞) | 402-208K | 113 | 0.68% | 1,296 | 7.76% | 1,385 | 8.29% |

Table 2. Read the fourth row as "terms in the '[2-4)' group were used 8 to 54 times. For the manual annotations there were 3,378 terms in this group, which is 20.2% of all GO terms." The percent columns sum to 100.

**Relationship between term characteristics and annotation usage**

While we are not surprised to observe a Zipfian distribution of term usage, the extreme skewness of the graphs in Figure 1 raises a broad question: is it possible to characterize terms that are used versus ones that are not? In an attempt to answer it, we test the following hypotheses: (1) there is a relationship between hierarchical depth and frequency of usage by annotators; (2) terms that are frequently used are

[d] We make the simplifying assumption that ln(0)=0.

less likely to be compositional than terms that are used infrequently. We first grouped the terms by their frequency of usage by creating five groups that correspond to even intervals on the log scale. These groups are given in Table 2. Figure 2 shows the relationship between frequency of usage and depth in the hierarchy. The average hierarchical depth of the terms in each group is shown. Statistically significant differences between these averages were determined using Kruskal-Wallis with a Bonferroni multiple comparison correction[e]. More frequently used terms tend to be higher in the hierarchy than infrequently used ones. Figure 3 shows the relationship between frequency of usage and compositionality. Terms that are more frequently used are less likely to have a contained term than terms that are less frequently used. We also examined the relationship between usage frequency and term length; the graph is similar to Figure 3 and is omitted for reasons of space. More frequently used terms tend to be shorter than infrequently used ones.

### 1.3. *Quantifying combinatorial effects*

A well-known difficulty of creating and maintaining a controlled terminology is the combinatorial explosion effects that present themselves when attempting to thoroughly represent all of the terms necessary to cover a domain [14]. Consider twenty of the terms that contain the words *T-cell* and *proliferation* shown in Figure 4. The blocked data highlights repeated data. It is apparent that the prefixes *regulation of*, *positive regulation of* and *negative regulation of* were added to five *T-cell proliferation* terms creating fifteen additional terms.

To better characterize this kind of combinatorial behavior we use a slightly modified finite state automaton (FSA) representation to represent a concise view of a set of GO terms[f]. Figure 5 represents all and only the 21 terms that contain the words *T-cell* and *proliferation*. No additional terms are represented. Any path that begins at a start state and ends in an end state corresponds to a subset of the graph's terms. A start state is represented by a single solid border, an end state is represented by double solid borders, and nodes with a dashed border are neither start states nor end states. The set of strings in a node represents a choice. A single GO

---

[e] All averages were found to differ with statistical significance (p=.05) except for the hierarchical depth averages for the groups labeled '0' and '(0-2).'

[f] FSAs are commonly used for representing regular expressions and grammars. We use rectangles instead of circles to reduce graph size.

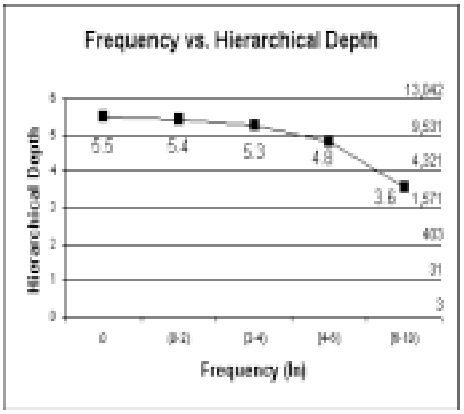


Figure 2. Frequency of term usage vs. hierarchical depth. Terms in the '[2-4)' group have an average hierarchical depth of 5.3. The cumulative number of terms at a given depth is shown on the right hand side of the graph, e.g. there are 1,571 terms at depths zero through three.

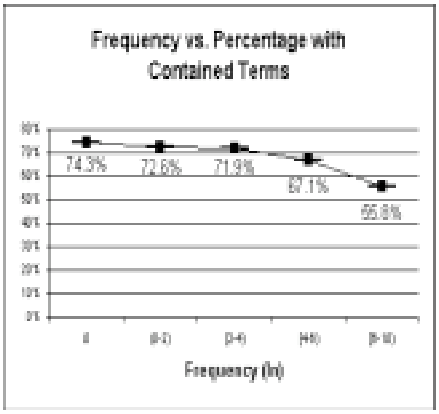


Figure 3. The y-axis shows the percentage of terms in a bin that have a contained term.

term is represented by a path for which a choice at each node has been made. The

| | | |
|---|---|---|
| | | T-cell proliferation |
| | | activated T-cell proliferation |
| | | alpha-beta T-cell proliferation |
| | | gamma-delta T-cell proliferation |
| | | T-cell homeostatic proliferation |
| | regulation of | T-cell proliferation |
| | regulation of | activated T-cell proliferation |
| | regulation of | alpha-beta T-cell proliferation |
| | regulation of | gamma-delta T-cell proliferation |
| | regulation of | T-cell homeostatic proliferation |
| positive | regulation of | T-cell proliferation |
| positive | regulation of | activated T-cell proliferation |
| positive | regulation of | alpha-beta T-cell proliferation |
| positive | regulation of | gamma-delta T-cell proliferation |
| positive | regulation of | T-cell homeostatic proliferation |
| negative | regulation of | T-cell proliferation |
| negative | regulation of | activated T-cell proliferation |
| negative | regulation of | alpha-beta T-cell proliferation |
| negative | regulation of | gamma-delta T-cell proliferation |
| negative | regulation of | T-cell homeostatic proliferation |

Figure 4. Twenty GO terms that contain *T-cell* and *proliferation*. The redundancy of data introduced by appending five terms with three modifiers is highlighted by the blocked text.

number of terms represented by a graph is given by:

$$\sum_{p \in G} \prod_{n \in p} |n|$$

This is a summation over the number of terms represented by each path (p) in the graph (G). The number of terms represented by a path is the product of the number of choices in each node (n) of the path. For example, the number of terms represented in Figure 5 is `(2*1*3*1*1) + (2*1*1*1) + … = 21`. We note that each term represented by this graph has a minimum edit distance of one from at least one other term in the graph.

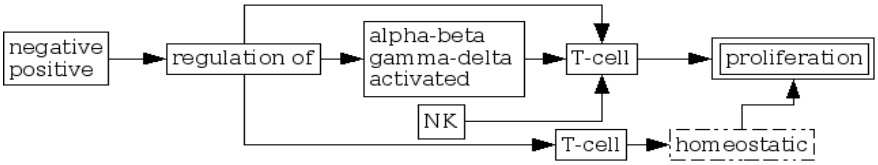


Figure 5. This FSA represents a concise view of all and only the 21 terms that contain the words *T-cell* and *proliferation*. The are two *T-cell* nodes because placing the *homeostatic* node between the other *T-cell* node and the *proliferation* node would license terms such as *alpha-beta T-cell homeostatic proliferation* that do not exist in GO.

This graphical representation is introduced only as a descriptive tool for visualizing terms in GO, not as a proposed solution for modeling GO. This representation, though inferior to other approaches for modeling GO, provides a very clear way to visualize and count terms and is sufficient for illustrating the combinatorial properties present in GO. We use this graphical representation to make two points: the number of terms licensed by simple compositional building blocks can be very large, and small conceptual changes to GO can result in large numbers of term changes.

A more complex example given in Figure 6 is a graphical term representation of all and only the 51 GO terms that contain the word *proliferation*. There are a number of observations about this set of terms that are easy to make when the 51 terms are displayed in this format. For instance, there are places where on could add an edge. For example, *regulation of neuroblast proliferation* is a reasonable GO term that could be represented by adding an edge from the *regulation of* node to the *neuroblast* node. Such observations would likely prove difficult to make by viewing a flat list of these 51 terms. Other edits that might be made to this graph include consolidating nodes and adding choices to individual nodes.

We next demonstrate that making small conceptual changes to the ontology results in a large number of terminological changes by observing the effect of making edits to the graph on the total number of terms represented by the graph. Consider the effect of making a conceptual change to the ontology such as "require differentiation and activation to appear in the same contexts as proliferation." To do this the strings *differentiation* and *activation* could be added to the two *proliferation* nodes in the graph. The effect of adding two choices to the rightmost node in the graph would nearly triple the number of terms represented by the graph to 151. To make an equivalent change to GO one must individually add or verify the existence of 100 terms. We also point out that each of these 100 terms differs by a single word from a proliferation term.

Coupling differentiation and activation to proliferation in this way is probably an undesirable oversimplification of the relationship between these three processes. We assessed the extent to which the terms that result from the kind of conceptual change described are valid. To do this we added the strings *differentiation* and *activation* to the rightmost node, creating 100 new terms. Of these 100 terms, 55 exist in GO. Of the 45 that do not, 14 or 31% of these were judged to be biologically meaningful and a reasonable addition to GO by two domain experts from the GO Consortium. Thus 69 of the 100 total "new" terms were biologically meaningful. This is evidence that there are relationships between proliferation, differentiation, and activation terms that require them to appear in similar contexts. It appears that the curation process for ensuring that these relationships are consistently applied when creating terms is error prone. In this case, reasonable terms were omitted. This may be due to the fact that the relationships themselves are not explicitly encoded or editable.

The graph in Figure 6 also suggests the possibility of simplifying the representation of *proliferation* terms by consolidating the various nodes that contain cell type descriptions by using a general *cell type* node as in Figure 7. The number of possible cell types that might fill in the *cell type* node is an open question. However, the OBO cell type ontology [15] contains 678 terms. Using this ontology in place of the *cell type* node along with the two additional choices in the rightmost node gives a graph that represents 8,136 terms! While not all of these 8,136 terms would be biologically meaningful, it would likely be difficult for a human curator to maintain a large subset of the 8,136 terms without some structure to manage the compositional building blocks that make up these terms.

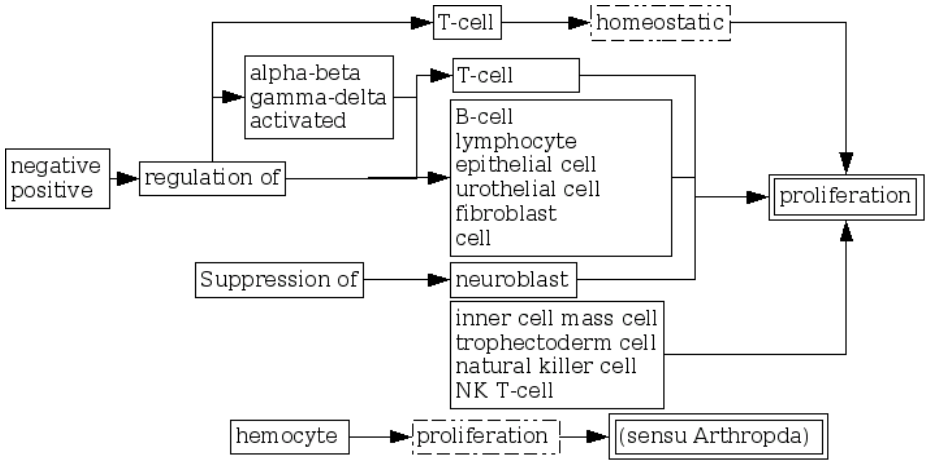


Figure 6. This FSA represents a concise view of all and only the 51 terms that contain the word *proliferation*.


Figure 7. Simplification of the finite state automaton shown in Figure 6.. The many cell types reflected in Figure 6 have been collapsed to a single *Cell Type* node.

The preceding examples from GO show that the possible number of terms can be incredibly large, and also demonstrate how such a collection of terms can be unwieldy to maintain. We note that we have found many other examples similar to the *proliferation* example that exhibit similar combinatorial behavior[g].

## 3. Discussion

Decades of research on relational databases have resulted in well-understood and widely accepted best practices for modeling relational data. Central to relational database design is the idea of table normalization. The key idea is to reduce the redundancy of data by thoroughly analyzing relationships between entities with respect to cardinality and directionality. Such analyses result in tables that have

[g] The terms containing *amino acid* provide a nice example because they are found in a different ontology (molecular function) and show that term variation may be at both the front and end of the terms.

reduced storage requirements and are much easier to populate and maintain consistently. We argue that the current representational model of the Gene Ontology is somewhat analogous to a database consisting of a single non-normalized table. Although GO's model is simple and easily understood, the terminological data is highly redundant and unnecessarily large due to the combinatorial effects demonstrated above. We argue that these characteristics have significantly impacted two activities vital to the success of the GO community: ontology curation and the use of GO for annotating databases. While we would not argue that reducing the terminological size is desirable or possible, it seems favorable to have a highly normalized model from which the terms are generated or derived. This gives curators and users two perspectives of GO. The first is a compositional model: well-normalized, compact, and complex. The second is the ontological model as it is: non-normalized and large, but easy to understand. The compositional model would presumably allow annotators to spend more time thinking about the relationships between terms rather than having to fill in the combinatorial possibilities that new or modified relationships license. For annotators, the compositional model would provide an alternate and smaller "search space" to navigate through when looking for terms.

The OBOL project attempts to address these issues in GO curation by creating a Prolog grammar to represent GO term compositionality. The grammar is proposed to be used to find missing relationships and terms in the ontology. In contrast, our approach focuses on the conceptual structure of GO, remaining neutral with respect to the representational format. We suggest that priority should be given to defining: (1) the proper compositional building blocks, (2) constraints that license their combinations, and (3) the domain relationships that parallel these combinations and constraints. Only after these aspects of the ontology are well understood do we think it makes sense to select a representational scheme (e.g. a rule-based grammar versus a frame-based system) on the basis of which best fits the conceptual structure of GO.

**Acknowledgments**

The authors thank Midori Harris, Jane Lomax, Chris Mungall, Sonia Leach, and Andrew Dolbey.

**References**


1. Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology." *Nature* 25:25-29 (2000).
2. Gene Ontology Consortium, "Creating the Gene Ontology resource: design and implementation." *Genome Research* 11:1425-1433 (2001).
3. C.M. Verspoor, C. Joslyn and G.J. Papcun, "The Gene Ontology as a source of lexical semantic knowledge for a biological natural language processing application," *Participant notebook of the ACM SIGIR'03 workshop on text analysis and search for bioinformatics,* pp. 51-56 (2003).
4. C.J. Wroe, R. Stevens, C.A. Goble and M. Ashburner, "A methodology to migrate the Gene Ontology to a description logic environment using DAML+OIL." *Pacific Symposium on Biocomputing 2003*.
5. I. Yeh, P.D. Karp, N.F. Noy and R.B. Altman, "Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO)." *Bioinformatics* 19(2):241-248, 2003.
6. C.A. Joslyn, S.M. Mniszewski, A. Fulmer and G. Heaton, "The Gene Ontology categorizer." *Bioinformatics,* in press.
7. C. Mungall, "Open Bio-Ontology Language." 7th Annual Bio-Ontologies Meeting (2004), http://bio-ontologies.man.ac.uk/obol-glasgow.ppt.
8. M. Bada, R. McEntire, C. Wroe and R. Stevens, "GOAT: The Gene Ontology Annotation Tool." *Proceedings of the 2003 UK e-Science All Hands Meeting*, 514-519 (2003).
9. K.M.Verspoor, J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha and T.Simas, "Protein annotation as term categorization in the Gene Ontology using word proximity networks." *BMC Bioinformatics* in submission.
10. P.V. Ogren, K.B. Cohen, G.K. Acquaah-Mensah, J. Eberlein, and L. Hunter, "The compositional structure of Gene Ontology terms." *Proceedings of the Pacific Symposium on Biocomputing 2004,* pp. 214-225.
11. D. Jurafsky and J.H. Martin, *Speech and language processing.* Prentice Hall (2000).
12. V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals." *Cybernetics and control theory* 10(8):707-710 (1966).
13. B.H. Partee, A. ter Meulen, and R.E. Wall, *Mathematical methods in linguistics,* corrected 1st edition. Kluwer Academic Publishers (1987).
14. A.L. Rector, "Clinical terminology: why is it so hard?" *Methods of Information in Medicine.* 38:239-52 (1999).
15. J. Bard and M. Ashburner, Cell Type Ontology, http://obo.sourceforge.net/ (2004).