A Lexical Semantic Network Induced from the Gene Ontology

Karin Verspoor, Cliff Joslyn, George Papcun
Los Alamos National Laboratory
PO Box 1663, MS B256
Los Alamos, NM 87545
+1 (505) 667-5086
{verspoor,joslyn,gjp}@lanl.gov

Keywords: Gene Ontology, lexical semantics, natural language processing

Abstract

We explore the Gene Ontology as a knowledge source for natural language processing applications in the biology domain. Using rules based on text parallelism, text insertion and modification relations applied to the hierarchical relations in the GO, we infer a network of lexical semantic relations implicit in the GO. The analysis of this network indicates that it contains significant information which can be used to augment the existing GO with the aim of constructing a broader knowledge resource for the biology domain, as well as to validate the relations in the GO.

1 Introduction

It is argued by Schulze-Kremer [7] that an ontology is a means to enable consistent use of concepts and terminology in different databases by providing a common set of concepts in which the semantic relations between the concepts are clearly specified. These properties of an ontology also make it attractive for use in a natural language processing (NLP) application, where there is a critical need to manage lexical (terminological) resources in a manner supporting representation of syntactic and semantic constraints on lexical use. In domains which contain much highly specific terminology, such as the biological domain, it is often a daunting task to construct such lexical resources. We turn, therefore, to existing terminological and ontological resources for the domain.

The Gene Ontology (GO, http://www.geneontology.org) [1] is a natural resource to consider using in the context of a NLP application in the biological domain. It is a sizeable, curated resource aiming to serve as a resource for consistent terminological use. It is, at core, a controlled vocabulary, but its real utility comes from the relationships specified between its vocabulary terms. The "is-a" and "part-of" hierarchies that exist in the GO not only contextualize individual terms, but provide a semantic grounding for those terms that can enable precise analysis of the meaning conveyed by those terms in relevant text sources. Despite identification of several difficulties with the GO as a formal ontological object [10], it is a valuable knowledge resource.

In this paper, we discuss explorations we have made into using the Gene Ontology as a source of lexical semantic knowledge for a text processing application in the biological

domain. The target use for the resulting lexicon is a prototype system, currently under development, that aims to extract regulatory relationships from biological text [6], and which depends on the existence of domain-specific lexical resources. While our customer has supplied some lists of terms that are associated with particular semantic types, these lists are invariably incomplete and exist independently of any domain ontology. We therefore look to the GO as a source of richer semantic data for lexical resources, specifically investigating its potential as a datasource enabling the incorporation of semantic generalizations into our NLP system. The work we present is complementary to the work of McCray et al [4] and Ogren et al [5], as we will discuss below, and is an elaboration of the work presented in [9].

2 Evidence of the Value of the GO for Text Processing

McCray et al [4] and Verspoor et al [9] explore the premise of using the GO in the context of a NLP application. McCray et al did an analysis of the occurrence of GO node labels (terms) with a large set of Medline abstracts (over 400,000), and found that 35% of the full GO node labels occurred in their corpus. The low percentage, 6%, of full GO node labels found by Verspoor et al in their, significantly smaller (9,336 Medline abstracts) corpus stands in stark constrast to this figure, and indicates the need for a sufficiently large corpus when searching for examples of terminology usage.

However, neither figure indicates high coverage of the corpus, precluding extensive *direct* use of the GO as a source of semantic information.[1] As such, Verspoor et al turned to an investigation of the overlap of individual words in GO node labels with the corpus. It was found that the words in the GO had good coverage of corpus words in the high- and middle-frequency ranges (above 63% for the high-frequency words, even in the small corpus investigated) indicating that for many of the terms we are likely to encounter regularly as we process domain texts, the GO may be able to provide a semantic grounding, if we are able to harness the semantics in the GO at the level of individual words rather than the full GO terms on nodes. This shift in focus to the level of individual words is further warranted based on McCray et al's results on the overlap of GO terms with the SPECIALIST lexicon in the Unified Medical Language System (UMLS) developed by the National Library of Medicine, showing only 9% of full GO terms occuring in that lexicon.

3 Inferring Lexical Relations

To find the implicit lexical relations which exist in the GO, we pursued a strategy of reasoning upon the ontological relations represented in the GO in order to establish ontological relations among smaller terms. Specifically, relations between heads of phrases are inferred from the relation between the phrases as a whole. We wish to evaluate the extent to which relations in the GO can be exploited in establishing relations between individual terms in the lexicon.

3.1 Inference Rules

Currently, our reasoning strategy for inducing lexical semantic relations from the GO utilizes three simple rules. These are not intended to capture the full range of lexical semantic

[1] Verspoor et al [9] define the direct use of the GO as directly utilizing the hierarchical relations in the GO for subsumption checking between individual (often multi-word) terms.

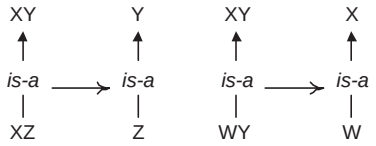

Figure 1: Text Parallelism Rule

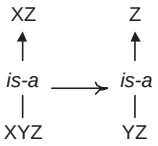

Figure 2: Insertion Rule

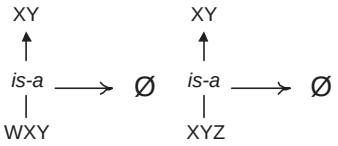

Figure 3: Modifier (Blocking) Rule

relations that might be induced from the GO, but rather are a first attempt in exploring whether there are meaningful relations that can be induced at all.

1. **Text Parallelism.** This rule attempts to infer an individual lexical relation from a recognized parallelism between phrases where there is some textual overlap between words. See Figure 1. For instance, from the GO relation "lipoprotein metabolism *isa* protein metabolism" we deduce "lipoprotein *isa* protein"; from "lipoprotein biosynthesis *isa* lipoprotein metabolism" we deduce "biosynthesis *isa* metabolism".

2. **Insertion.** This rule handles the case in which a word (or words) are inserted in the middle of a term, creating a child term as a specialization of a parent term. See Figure 2. We have implemented the rule to allow grouping to the right, based on the right-branching structure of English. While this grouping will not always reflect the most intuitive structure of a phrase, in the context of the GO this seems to be more common than a left-branching structure and without implementing full parsing we need to make a (somewhat arbitrary) choice. When this rule is applied, the GO relation "adult feeding behavior *isa* adult behavior" results in the inference "feeding behavior *isa* behavior"; from "chemosensory jump behavior *isa* chemosensory behavior" we deduce "jump behavior *isa* behavior"

3. **Modifier.** This rule handles the case in which one term is a specialization of the other through the introduction of a pre- or post-modifier. See Figure 3.

   In this case, the rule disallows an inference, following from the recognition that the the modifiers generally modify the entire phrase, and any relation at the level of individual lexical item doesn't make sense. For instance, there is no clear lexical relation to be inferred from "positive gravitactic behavior *isa* gravitactic behavior" or "larval feeding behavior (sensu insecta) *isa* larval feeding behavior". Inferring "positive gravitactic *isa* gravitactic" or "behavior (sensu insecta) *isa* behavior" would not accurately capture the semantics of the original relation. However, it is not the case that the observation of these modification relations in the GO is not useful for analysis of the GO; in fact it is precisely these relations which Ogren et al [5] draw on to infer semantic constraints on what they call *derivational phrases*. We will discuss this in Section 5.

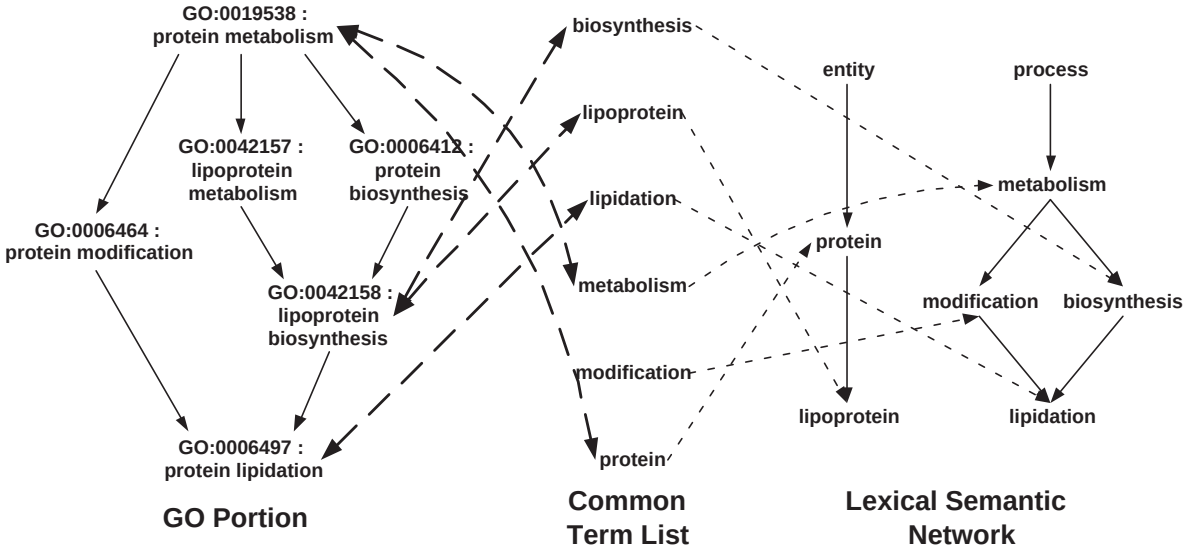

Figure 4: Mappings from the GO to a lexical semantic network

3.2 Relations Inferred

These rules can be applied to each parent-child pair in the GO, giving us a set of additional parent-child pairs that can be integrated to form a lexical semantic network. Figure 4 illustrates how these lexical semantic inferences can be linked via the terms they involve to the GO.

In applying these rules, we generated additional parent-child pairs for 9,574 out of 16,849 parent-child pairs in the GO (57%).[2] This corresponds to 6,364 unique parent-child relations. The top-ranking relations are shown in Table 1, along with the number of times the relation was inferred. We believe that these reflect some fundamental relations; these can form the starting point for a domain ontology at the lexical level as well as the phrasal level. Some of these relations do correspond to existing parent-child pairs in the GO (such as the first two in the table, which correspond to generic physiological processes), but others do not, such as the relationship between RNA, and tRNA, mRNA, rRNA, and snRNA. Overall, only 70 of the 6,364 generated relations already existed in the GO; in Table 1 all but the first two are new parent-child relations not found in the GO.

Of the 6,589 unique node labels in the set of parent-child relations induced, nearly half (3,270) do not exist in the GO as node labels (there are 12,881 unique node labels in the set of original parent-child relations from the GO that we worked with). This indicates clearly that we have generated relations involving entities not provided any explicit semantic grounding in the GO. From these figures, we can conclude that the relations inferred through our reasoning process capture information that does not exist explicitly in the GO, and therefore is potentially valuable new information.

[2]Note that the figures we report in this paper are different from the results reported in [9] due to an error in the scripts used for the previous results.

Table 1: Lexical semantic relations induced from GO

581	biosynthesis isa metabolism
577	catabolism isa metabolism
44	receptor isa binding
38	deoxyribonucleoside isa nucleoside
35	ribonucleoside isa nucleoside
33	permease isa transporter
27	Saccharomyces isa Fungi
22	porter isa transporter
15	oxidation isa metabolism
14	tRNA isa RNA
14	inhibitor isa regulator
13	ribonucleotide isa nucleotide
11	proliferation isa activation
11	differentiation isa activation
11	deoxyribonucleotide isa nucleotide
10	rRNA isa RNA
10	mRNA isa RNA
9	snRNA isa RNA
8	modification isa metabolism
8	methylation isa modification

3.3 A network of relations

We combined the inferred parent-child relationships into a network, in order to get a sense of the structure of the graph resulting from the inference process. Using the Pajek network visualization tool [2], we found the network too large to allow for significant manual analysis. However, with Pajek we were able to reduce the full network to a more manageable hierarchy. While this reduction loses some information (for instance, cycles and nodes with multiple parents are removed), it helped us to explore basic properties of the graph structure. The generated hierarchy consists of a forest of trees, and contains 5,447 of the original 6,364 relations. The most salient property of this hierarchy is that the vast majority of the inferred relations do not embed within other inferred relations. There are only 391 nodes which are both a parent and a child. As such, there are many trees of length 2. Specifically, there are 773 trees in the generated hierarchy, of which 669 have length 2 and 69 length 3. The root covering the largest number of nodes is "activity", with 1,149 descendants at a maximum depth of 4. The deepest tree in the hierarchy has depth 10, rooted at "biosynthesis". This can be compared to the structure of the GO itself, which Joslyn et al [3] reports to have a maximum chain length of 16, including the top "Gene Ontology" root node. The network we have inferred is significantly flatter than the GO, as might be expected given that the terms involved are structurally simpler (and shorter) than those occurring in the GO, since they have been derived from portions of the original GO terms.

Figure 5 shows the tree in the induced lexical semantic hierarchy with depth 9, the second deepest tree in the collection. This tree gives a sense of the upper bound of the ontological complexity that can be expected through application of our current reasoning

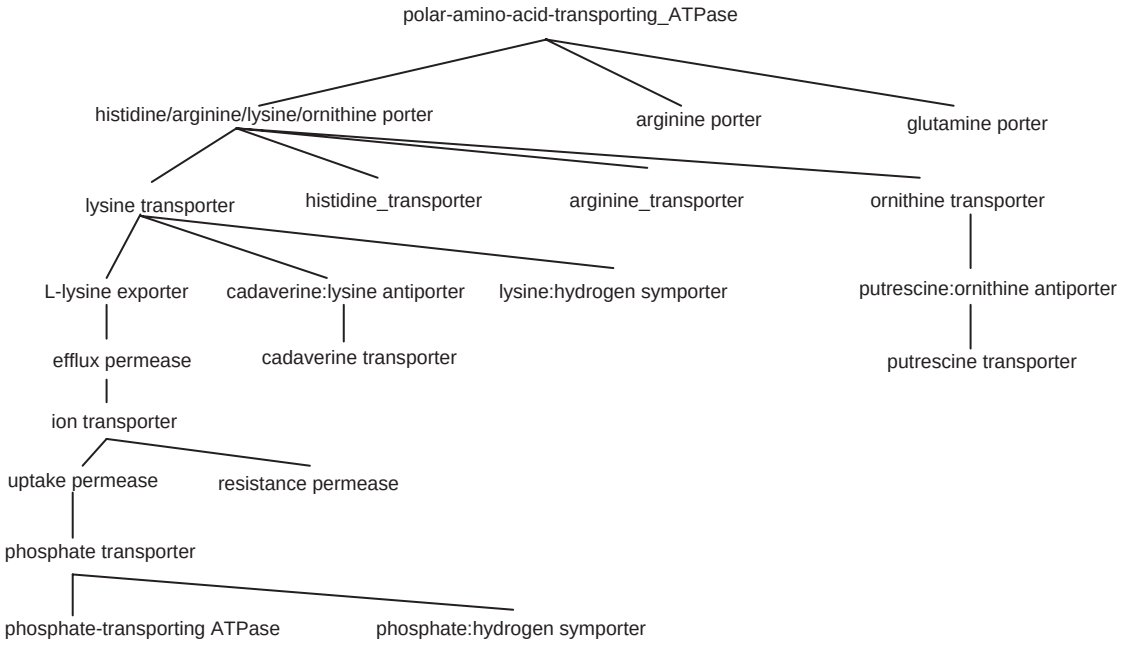

Figure 5: The second deepest tree in the induced hierarchy

Table 2: The children of "enzyme"

transposase	transferase	ribonuclease
proteasome	protease	phospholipase
phosphatase	pectinesterase	oxidoreductase
MAPK	lyase	7,8-dihydro-8-oxoguanine-triphosphatase
kinase	isomerase	integrase
hydrolase	helicase	GTPase
diazepam-binding	cyclase	chaperone
caspase	ATPase	ATPase stimulator activator
alpha-amylase	aromatase	ligase

strategy. Table 2 shows the children rooted at "enzyme", which is a flat tree. This list was given to a domain expert, who rejected "diazepam-binding" as a valid child of "enzyme",[3] questioned the inclusion of "proteasome", and accepted the remaining 25. So while the results of the inferences are not perfect, and the full set of inferences would need to be validated before integration with the GO itself, this example does show the potential value of the approach: we have identified a concept, "enzyme", which does not exist in the GO as an individual node, and we have been able to provide it with some semantic grounding, specifically a (partial) listing of the kinds of entities that are enzymes. Even such flat trees can be extremely valuable in supporting generalizations, in particular if a concept relates many entities, as in this case.

[3]This relation was inferred from the GO relation "diazepam-binding inhibitor activity *isa* enzyme inhibitor activity". If the inference is invalid, we might have to question the source GO relation as well, at least to understand the intended interpretation of the relation.

Table 3: Relations in the GO leading to cycles

binding *isa* **transporter**
lipopolysaccharide binding activity *isa* lipopolysaccharide transporter activity
transporter *isa* **binding**
sphingolipid transporter activity *isa* sphingolipid binding activity
oxygen transporter activity *isa* oxygen binding activity
phosphatidylinositol transporter activity *isa* phosphatidylinositol binding activity
phospholipid transporter activity *isa* phospholipid binding activity
modification *isa* **processing**
RNA modification *isa* RNA processing
mRNA modification *isa* mRNA processing
processing *isa* **modification**
protein processing *isa* protein modification

3.4 Analysis of the Inferences

The rules also result in some problematic inferences. For instance, the right-branching preference in the Insertion rule when applied to "adult male behavior *isa* adult behavior" results in the inference "male behavior *isa* behavior". This inference is not incorrect, but intuitively one would prefer the inference of "adult male *isa* male" from this source relation. This could perhaps be modeled through the incorporation of statistical parsing or, more straightforwardly, reference to the relative mutual information of the alternative phrasal analyses. We have not yet tried this.

The Parallel rule sometimes leads to inferences that, independently, seem quite odd. For instance, application of the rule to "maternal behavior *isa* reproductive behavior" and "mating behavior *isa* reproductive behavior" results in "maternal *isa* reproductive" and "mating *isa* reproductive". The inferred relations are rather forced and difficult to interpret. What seems to be going on in this case is (a) there is a context-dependent interpretation of the relationship between the adjective and the noun in these two phrases which is lost when the nominal context is removed (where the parent/child relation expresses something like "maternal behavior" *isa* "behavior in support of successful reproduction") and (b) the *isa* relation does not adequately capture the relation between the parent and the child – in what sense is a maternal behavior *really* a reproductive behavior?

The lexical semantic network which is generated via these rules from the GO can be used to augment the GO itself, in order to extend the GO from a collection of phrasal relations to a more detailed ontology. Along the way, this approach will help to validate the information in the GO by highlighting instances where the *isa* relation may be insufficient, or even by identifying cases where there might be inconsistencies in the GO through recognition of a cycle in the lexical semantic network.

We did in fact find two such cycles in our inferred network. Looking for 2-cycles only, we found both "transporter *isa* binding" and "binding *isa* transporter", as well as "modification *isa* processing" and "processing *isa* modification". Tracing these back to the source relations of these inferences in the GO, we find the relations as indicated in Table 3. Clearly there is some inconsistency here in the use of these terms, if not outright errors in the GO itself.

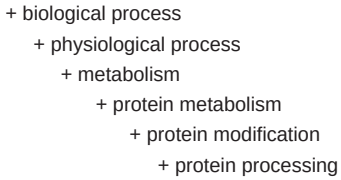

Figure 6: Path to "protein processing" in the GO

It would seem, for instance, that based on the generic sense in English of "processing", "protein processing" should be more abstract than "protein modification" and that therefore the relation should exist in the opposite direction in the GO. However, looking at the path to this relation in the GO as represented in Figure 6, it is possible that "processing" in this instance is intended to refer to a specific type of protein modification distinguished in some important, but unspecified, way from the other kinds of protein modification. The problem is that if the word is being used in a way which violates our natural expectations, this usage should be specifically defined and differentiated from the expected meaning. This observation is in line with Schulze-Kremer [7]'s proposal to explicitly specify the criteria for subclassifying concepts in an ontology. The GO does not even provide a definition for the "protein processing" term; what the cycle we have found indicates is that there is some ambiguity in the use of the word "processing" which needs to be resolved.

4 Application

Ultimately, our goal is to incorporate these lexical relations into a NLP system which aims to extract regulatory relationships from Medline abstracts [6]. In the lexicon used by the NLP system, we can define mappings of ontological categories from GO to lexical items. With this in place, lexical items can be considered in the far richer semantic context provided by the GO. This is achieved by incorporating subsumption checking into the patterns which drive the information extraction. For instance, a rule may require that a particular argument be some type of protein metabolism. With reference to the GO, and the additional lexical semantic relations we have induced, we can verify that this holds for a given word or phrase identified in the text. These types of constraints allow us to more accurately identify particular relationships.

As an example, in our NLP system we may wish to identify all sentences in which a protein is acting metabolically. Rather than spelling out all the different kinds of metabolic function, we can draw on the structure of the ontology. For instance, we might define a pattern [PROTEIN serves a METABOLISM function], where we verify that the word preceding *function* maps to a node in the ontology subsumed by *metabolism*. The term *biosynthetic*, for example, maps to the ontology node *biosynthesis*, that is in turn subsumed by *metabolism*. So the sentence "The lipoprotein serves a biosynthetic function" could be identified as satisfying the more general pattern, although it mentions a lipoprotein rather than a protein, and biosynthesis rather than metabolism.

5 Validating and Augmenting the GO

As we have seen, the application of the simple inference rules to the GO results in a network which facilitates validation of the GO itself. Cycles in the inferred network indicate an

inconsistency in the usage of a term; questionable or invalid relations suggest the need for examination of the source relation in the GO.

Some of the questionable inferred relations could of course be due to the simplicity of the inference rules; as suggested above, the right-grouping heuristic of the insertion rule requires refinement with more sophisticated structural analysis of the GO terms. Similarly, there are conventions in the structure of GO terms that are not taken into consideration in the inference rules, such as the use of a colon (":") to indicate molecules interacting in a reaction in specific ways. The rules should be refined to take such conventions into consideration; doing so requires the semantics of the conventions to be clear.

The approach we have outlined is based on examining local, parent-child relations in the GO. Ogren et al [5], in contrast, search for systematic repetitions of phrases (substrings) across all GO terms, not just locally in individual relations, in order to identify phrases encoding specific semantic relations. These *derivational phrases* can provide two types of semantic groundings for concepts – the phrases can have a specific, consistent meaning that should be explicitly characterized, and the phrases can impose semantic constraints on the terms they modify. Ogren et al show that such constraints can be inferred by the positional distribution of the modified terms. This type of inference cannot be made locally, and is likely to be extremely useful in the context of natural language processing, which often depends on semantic constraints to resolve ambiguities.

Ogren et al [5] show how identification of derivational phrases can help uncover terms which are candidates to be GO terms, but are not. They identified 108 candidate terms on this basis. Our approach identified 3,270 new candidate terms, as well as providing each of these terms with at least one ontological relation as a starting point for providing the appropriate semantic grounding of the terms. However, many of our candidate terms are of a different nature than the terms currently in the GO – concepts like "enzyme" and "RNA" are general biological concepts and not concepts primarily extending from biological process, molecular function, or cellular composition. It may not be appropriate to consider adding these to the GO itself – but identifying those concepts and relations that they participate in is useful for meeting the need of constructing general knowledge resources for the biology domain. Such a resource would be valuable not only in support of natural language processing applications in the biology domain, but also to provide tools to assist in the validation of the GO (through identification of inconsistencies in term usage) and the automatic augmentation of the GO (e.g. to support abstraction of existing or new terms on the basis of known semantic relations, as suggested by Williams and Andersen [10]).

6 Conclusion and Future Work

In this work we have investigated the potential for exploiting the Gene Ontology, an ontology in the biology domain, as a source of the kind of lexical semantic knowledge. We have shown that the application of some simple inference rules to the parent/child pairs in the GO can result in the creation of a semantic network that captures core lexical relations for the domain, and can be used to enable generalization in our information extraction system. The GO itself could be augmented, and in turn validated, with these lexical relations.

There are several tasks that remain as future work, including more formal evaluation of the induced lexical semantic network by domain experts and further analysis of the structure

of the network, including the nodes with multiple parents that were stripped out through the network reduction we performed. We would like to refine the inference rules used, and we would like to try an incremental reasoning strategy where we apply the inference rules to the induced network, to see what second-order inferences can be drawn. We would also like to try integrating the lexical semantic network with the GO itself, to further explore the possibility of cycles and inconsistencies which might be exposed in so doing. Finally, we wish to explore the integration of the local semantic relations we have inferred with the globally-derived inferences about semantic constraints and semantic function as reported by Ogren et al [5]. The integration of these two complementary views of the lexical semantics of the Gene Ontology may prove to be the most effective way of exploiting all of the knowledge existing implicitly in the GO.

7 Acknowledgments

This work was supported through a Los Alamos National Laboratory collaboration with Procter & Gamble Corporation, and by the Department of Energy under contract W-7405-ENG-36 to the University of California.

References

[1] Ashburner, M., C. Ball, and J. B. et al. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

[2] Batagelj, V., A. Mrvar. 2003. Pajek - Analysis and Visualization of Large Networks. In Juenger, M., Mutzel, P. (Eds.): Graph Drawing Software. Springer, Berlin, pp. 77-103.

[3] Joslyn, C., S. Mniszewski, A. Fulmer, G. Heaton. 2004. The Gene Ontology Categorizer. Submitted to ISMB 2004.

[4] McCray, A., A. Browne, and O. Bodenreider. 2002. The Lexical Properties of the Gene Ontology (GO). In *Proceedings of the AMIA 2002 Annual Symposium*, pp. 504–508.

[5] Ogren, P., K. B. Cohen, G. Acquaah-Mensah, J. Eberlein, and L. Hunter. 2004. The compositional structure of Gene Ontology terms. In *Proceedings of the Pacific Symposium on Biocomputing*.

[6] Papcun, G., K. Sentz, A. Fulmer, J. Xu, O. Lubeck, and M. Wolinsky. 2003. A construction grammar approach to extracting regulatory relationships from biological literature. In *Proceedings of the Pacific Symposium on Biocomputing*, Kauai, Hawaii.

[7] Schulze-Kremer, S. 1998. Ontologies for molecular biology. In *Pacific Symposium on Biocomputing*, AAAI Press: 695-706.

[8] Verspoor, C. M., C. Joslyn, and G. Papcun. 2003. Interactions between the gene ontology and a domain corpus for a biological natural language processing application. In *Proceedings of the ISMB'03 Workshop on Bioontologies*, Brisbane, Australia.

[9] Verspoor, C. M., C. Joslyn, and G. Papcun. 2003. The Gene Ontology as a Source of Lexical Semantic Knowledge for a Biological Natural Language Processing Application. In *Proceedings of the SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics*, Toronto, CA, August 1.

[10] Williams, J. and W. Andersen. 2003. Bringing Ontology to the Gene Ontology *Comp Funct Genom*; 4:90-93.