Integrating Ontological Knowledge and Textual Evidence in Estimating Gene and Gene Product Similarity

Antonio Sanfilippo, Christian Posse, Banu Gopalan, Stephen Tratz, Michelle Gregory

Pacific Northwest National Laboratory
Richland, WA 99352
{Antonio.Sanfilippo, Christian.Posse, Banu.Gopalan, Stephen.Tratz, Michelle.Gregory}@pnl.gov

Abstract

With the rising influence of the Gene Ontology, new approaches have emerged where the similarity between genes or gene products is obtained by comparing Gene Ontology code annotations associated with them. So far, these approaches have solely relied on the knowledge encoded in the Gene Ontology and the gene annotations associated with the Gene Ontology database. The goal of this paper is to demonstrate that improvements to these approaches can be obtained by integrating textual evidence extracted from relevant biomedical literature.

1 Introduction

The establishment of similarity between genes and gene products through homology searches has become an important discovery procedure that biologists use to infer structural and functional properties of genes and gene products–see Chang et al. (2001) and references therein. With the rising influence of the Gene Ontology[1] (GO), new approaches have emerged where the similarity between genes or gene products is obtained by comparing GO code annotations associated with them. The Gene Ontology provides three orthogonal networks of functional genomic concepts structured in terms of semantic relationships such as inheritance and meronymy, which encode biological process (BP), molecular function (MF) and cellular component (CC) properties of genes and gene products. GO code annotations explicitly relate genes and gene products in terms of participation in the same/similar biological processes, presence in the same/similar cellular components and expression of the same/similar molecular functions. Therefore, the use of GO code annotations in establishing gene and gene product similarity provides significant added functionality to methods such as BLAST (Altschul et al. 1997) and FASTA (Pearson and Lipman 1988) where gene and gene product similarity is calculated using string-based heuristics to select maximal segment pair alignments across gene and gene product sequences to approximate the Smith-Waterman algorithm (Smith and Waterman 1981).

Three main GO-based approaches have emerged so far to compute gene and gene product similarity. One approach assesses GO code similarity in terms of shared hierarchical relations within each gene ontology (BP, MF, or CC) (Lord et al. 2002, 2003; Couto et al. 2003; Azuaje et al. 2005). For example, the relative semantic closeness of two biological processes would be determined by the informational specificity of the most immediate parent that the two biological processes share in the BP ontology. The second approach establishes GO code similarity by leveraging associative relations across the three gene ontologies (Bodenreider et al. 2005). Such associative relations make predictions such as which cellular component is most likely to be the location of a given biological proc-

[1] http://www.geneontology.org.

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 25–32,
New York City, June 2006.

ess and which molecular function is most likely to be involved in a given biological process. The third approach computes GO code similarity by combining hierarchical and associative relations (Posse et al. 2006).

Several studies within the last few years (Andrade et al. 1997, Andrade 1999, MacCallum et al. 2000, Chang at al. 2001) have shown that the inclusion of evidence from relevant scientific literature improves homology search. It is therefore highly plausible that literature evidence can also help improve GO-based approaches to gene and gene product similarity. Sanfilippo et al. (2004) propose a method for integrating literature evidence within an early version of the GO-based similarity algorithm presented in Posse et al. (2006). However, no effort has been made so far in evaluating the potential contribution of textual evidence extracted from relevant biomedical literature for GO-based approaches to the computation of gene and gene product similarity. The goal of this paper is to address this gap with specific reference to the assessment of protein similarity.

2 Background

GO-based similarity methods that focus on measuring intra-ontological relations have adopted the information theoretic treatment of semantic similarity developed in Natural Language Processing−see Budanitsky (1999) for an extensive survey. An example of such a treatment is given by Resnik (1995), who defines semantic similarity between two concept nodes *c1 c2* in a graph as the information content of the least common superordinate (*lcs*) of *c1* and *c2*, as shown in (1). The information content of a concept node *c, IC(c),* is computed as *-log p(c)* where p*(c)* indicates the probability of encountering instances of *c* in a specific corpus.

$$(1)\quad sim(c1,c2) = IC(lcs(c1,c2)) = -\log p(lcs(c1,c2))$$

Jiang and Conrath (1997) provide a refinement of Resnik's measure by factoring in the distance from each concept to the least common superordinate, as shown in (2).[2]

$$(2)\quad sim(c1,c2) = \frac{1}{IC(c1)+IC(c2)-2\times IC(lcs(c1,c2))}$$

Lin (1998) provides a slight variant of Jiang's and Conrath's measure, as indicated in (3).

$$(3)\quad sim(c1,c2) = \frac{2 \times IC(lcs(c1, c2))}{IC(c1) + IC(c2)}$$

The information theoretic approach is very well suited to assess GO code similarity since each gene subontology is formalized as a directed acyclic graph. In addition, the GO database[3] includes numerous curated GO annotations which can be used to calculate the information content of each GO code with high reliability. Evaluations of this methodology have yielded promising results. For example, Lord et al. (2002, 2003) demonstrate that there is strong correlation between GO-based similarity judgments for human proteins and similarity judgments obtained through BLAST searches for the same proteins. Azuaje et al. (2005) show that there is a strong connection between the degree of GO-based similarity and the expression correlation of gene products.

As Bodenreider et al. (2005) remark, the main problem with the information theoretic approach to GO code similarity is that it does not take into account associative relations across the gene ontologies. For example, the two GO codes 0050909 (*sensory perception of taste*) and 0008527 (*taste receptor activity*) belong to different gene ontologies (BP and MF), but they are undeniably very closely related. The information theoretic approach would simply miss associations of this kind as it is not designed to capture inter-ontological relations.

Bodenreider et al. (2005) propose to recover associative relations across the gene ontologies using a variety of statistical techniques which estimate the similarity of two GO codes inter-ontologically in terms of the distribution of the gene product annotations associated with the two GO codes in the GO database. One such technique is an adaptation of the vector space model frequently used in Information Retrieval (Salton et al. 1975), where

[2] Jiang and Conrath (1997) actually define the distance between two concepts nodes c1 c2, e.g.

$dist(c1, c2) = IC(c1) + IC(c2) - 2 \times IC(lcs(c1, c2))$

For ease of exposition, we have converted Jiang's and Conrath's semantic distance measure to semantic similarity by taking its inverse, following Pedersen et al. (2005).

[3] http://www.godatabase.org/dev/database.

each GO code is represented as a vector of gene-based features weighted according to their distribution in the GO annotation database, and the similarity between two GO codes is computed as the cosine of the vectors for the two codes.

The ability to measure associative relations across the gene ontologies can significantly augment the functionality of the information theoretic approach so as to provide a more comprehensive assessment of gene and gene product similarity. However, in spite of their complementarities, the two GO code similarity measures are not easily integrated. This is because the two measures are obtained through different methods, express distinct senses of similarity (i.e. intra- and inter-ontological) and are thus incomparable.

Posse et al. (2006) develop a GO-based similarity algorithm–XOA, short for Cross-Ontological Analytics–capable of combining intra- and inter-ontological relations by "translating" each associative relation across the gene ontologies into a hierarchical relation within a single ontology. More precisely, let *c1* denote a GO code in the gene ontology *O1* and *c2* a GO code in the gene ontology *O2*. The XOA similarity between *c1* and *c2* is defined as shown in (4), where[4]

- *cos(ci,cj)* denotes the cosine associative measure proposed by Bodenreider et al. (2005)
- *sim(ci,cj)* denotes any of the three intra-ontological semantic similarities described above, see (1)-(3)
- $\max_{ci \text{ in } Oj} \{f(ci)\}$ denotes the maximum of the function f() over all GO codes *ci* in the gene ontology *Oj*.

The major innovation of the XOA approach is to allow the comparison of two nodes *c1*, *c2* across distinct ontologies *O1, O2* by mapping *c1* into its closest node *c4* in *O2* and *c2* into its closest node *c3* in *O1*. The inter-ontological semantic similarity between *c1* and *c2* can be then estimated from the intra-ontological semantic similarities between *c1-c3* and *c2-c4*, using multiplication with the associative relations between *c2-c3* and *c1-c4* as a score enrichment device.

$$(4)\quad \mathrm{XOA}(c1, c2) = \max\left\{ \begin{array}{l} \max_{c3 \text{ in } O1} \left\{ \begin{array}{l} \mathrm{sim}(c1, c3)\times \\ \cos(c2, c3) \end{array} \right\}, \\ \\ \max_{c4 \text{ in } O2} \left\{ \begin{array}{l} \mathrm{sim}(c2, c4)\times \\ \cos(c1, c4) \end{array} \right\} \end{array} \right\}$$

Posse et al. (2006) show that the XOA similarity measure provides substantial advantages. For example, a comparative evaluation of protein similarity, following the benchmark study of Lord et al. (2002, 2003), reveals that XOA provides the basis for a better correlation with protein sequence similarities as measured by BLAST bit score than any intra-ontological semantic similarity measure. The XOA similarity between genes/gene products derives from the XOA similarity between GO codes. Let *GP1* and *GP2* be two genes/gene products. Let *c11*,*c12*,…, *c1n* denote the set of GO codes associated with *GP1* and *c21*, *c22*,…., *c2m* the set of GO codes associated with *GP2*. The XOA similarity between *GP1* and *GP2* is defined as in (5), where *i=1,…,n* and *j=1,…,m*.

$$(5)\ \mathrm{XOA}(GP1,GP2) = \max\ \{\mathrm{XOA}(c1i, c2j)\}$$

The results of the study by Posse et al. (2006) are shown in Table 1. Note that the correlation between protein similarities based on intra-ontological similarity measures and BLAST bit scores in Table 1 is given for each choice of gene ontology (MF, BP, CC). This is because intra-ontological similarity methods only take into account GO codes that are in the same ontology and can therefore only assess protein similarity from a single ontology viewpoint. By contrast, the XOA-based protein similarity measure makes use of GO codes that can belong to any of the three gene ontologies and needs not be broken down by single ontologies, although the contribution of each gene ontology or even single GO codes can still be fleshed out, if so desired.

Is it possible to improve on these XOA results by factoring in textual evidence? We will address this question in the remaining part of the paper.

[4] If *c1* and *c2* are in the same ontology, i.e. *O1=O2*, then *xoa(c1,c2)* is still computed as in (4). In most cases, the maximum in (4) would be obtained with *c3* = *c2* and *c4* = *c1* so that XOA*(c1,c2)* would simply be computed as *sim(c1,c2)*. However, there are situations where there exists a GO code *c3* (*c4*) in the same ontology which

- is highly associated with *c1* (*c2*),
- is semantically close to *c2* (*c1*), and
- leads to a value for *sim(c1,c3)* x *cos(c2,c3)* *((sim(c2,c4)* x *cos(c1,c4))* that is higher than *sim(c1,c2)*.

Semantic Similarity Measures		Resnik	Lin	Jiang & Conrath
Intra-ontological				
	Molecular Function	0.307	0.301	0.296
	Biological Process	0.195	0.202	0.203
	Cellular Component	0.229	0.234	0.233
XOA		**0.405**	**0.393**	**0.368**

Table 1: Spearman rank order correlation coefficients between BLAST bit score and semantic similarities, calculated using a set of 255,502 protein pairs–adapted from Posse et al. (2006).

3 Textual Evidence Selection

Our first step in integrating textual evidence into the XOA algorithm is to select salient information from biomedical literature germane to the problem. Several approaches can be used to carry out this prerequisite. For example, one possibility is to collect documents relevant to the task at hand, e.g. through PubMed queries, and use feature weighting and selection techniques from the Information Retrieval literature−e.g. *tf*idf* (Buckley 1985) and Information Gain (e.g. Yang and Pedersen 1997)−to distill the most relevant information. Another possibility is to use Information Extraction algorithms tailored to the biomedical domain such as *Medstract* (http://www.medstract.org, Pustejovsky et al. 2002) to extract entity-relationship structures of relevance. Yet another possibility is to use specialized tools such as GoPubMed (Doms and Schroeder 2005) where traditional keyword-based capabilities are coupled with term extraction and ontological annotation techniques.

In our study, we opted for the latter solution, using generic Information Retrieval techniques to normalize and weigh the textual evidence extracted. The main advantage of this choice is that tools such as GoPubMed provide very high quality term extraction at no cost. Less appealing is the fact that the textual evidence provided is GO-based and therefore does not offer information which is orthogonal to the gene ontology. It is reasonable to expect better results than those reported in this paper if more GO-independent textual evidence were brought to bear. We are currently working on using *Medstract* as a source of additional textual evidence.

GoPubMed is a web server which allows users to explore PubMed search results using the Gene Ontology for categorization and navigation purposes (available at http://www.gopubmed.org). As shown in Figure 1 below, the system offers the following functionality:

- It provides an overview of PubMed search results by categorizing abstracts according to the Gene Ontology
- It verifies its classification by providing an accuracy percentage for each
- It shows definitions of Gene Ontology terms
- It allows users to navigate PubMed search results by GO categories
- It automatically shows GO terms related to the original query for each result
- It shows query terms (e.g. “Rab5” in the middle windowpane of Figure 1)
- It automatically extracts terms from search results which map to GO categories (e.g. highlighted terms other than “Rab5” in the middle windowpane of Figure 1).

In integrating textual evidence with the XOA algorithm, we utilized the last functionality (automatic extraction of terms) as an Information Extraction capability. Details about the term extraction algorithm used in GoPubMed are given in Delfs et al. (2004). In short, the GoPubMed term extraction algorithm uses word alignment strategies in combination with stemming to match word sequences from PubMed abstracts with GO terms. In doing so, partial and discontinuous matches are allowed. Partial and discontinuous matches are weighted according to closeness of fit. This is indicated by the accuracy percentages associated with GO in Figure 1 (right side). In this study we did not make use of these accuracy percentages, but plan to do so in the future.

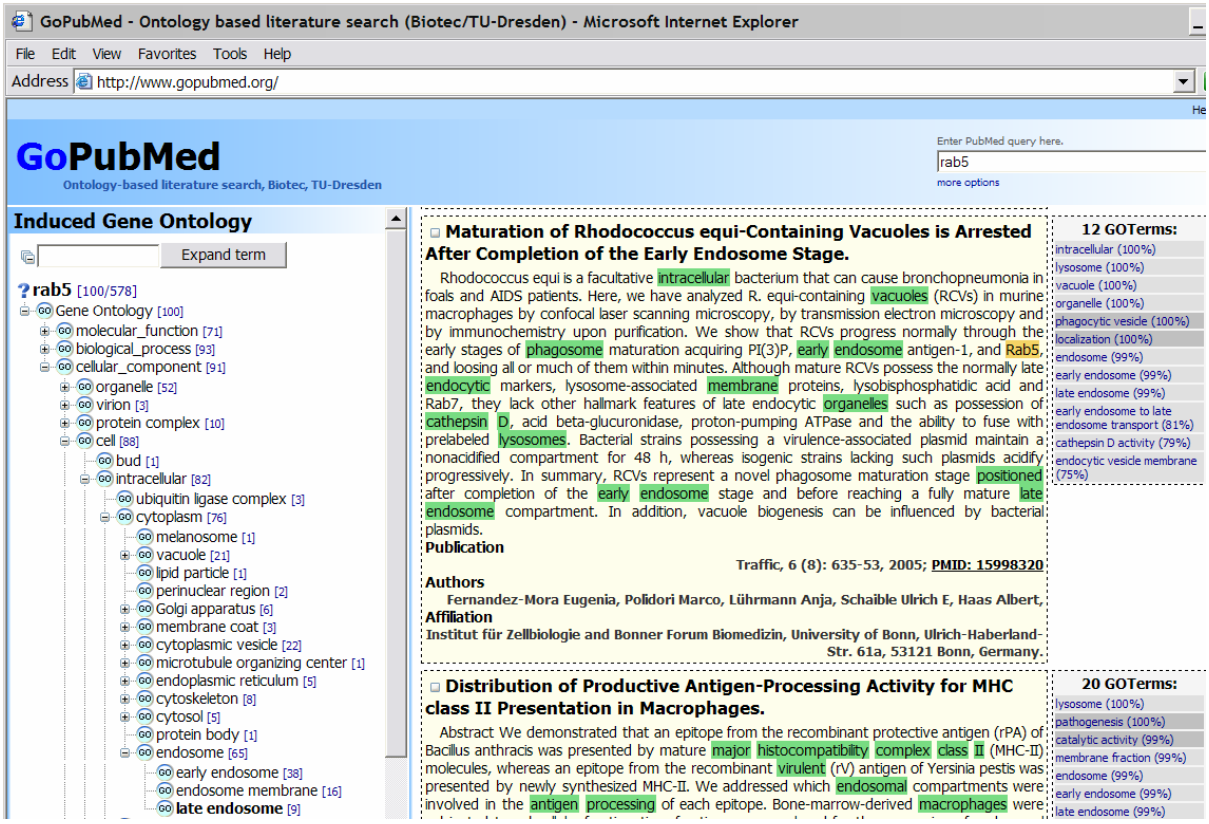

Figure 1: GoPubMed sample query for the "rab5" protein. The abstracts shown are automatically proposed by the system after the user issues the protein query and then selects the GO term "**late endosome**" (bottom left) as the discriminating parameter.

Our data set consists of 2360 human protein pairs containing 1783 distinct human proteins. This data set was obtained as a 1% random sample of the human proteins used in the benchmark study of Posse et al. (2006)–see Table 1.[5] For each of the 1783 human proteins, we made a GoPubMed query and retrieved up to 100 abstracts. We then collected all the terms extracted by GoPubMed for each protein across the abstracts retrieved. Table 2 provides an example of the output of this process.

nutrient, uptake, carbohydrate, metabolism, affecting, cathepsin, activity, protein, lipid, growth, rate, habitually, signal, transduction, fat, protein, cadherin, chromosomal, responses, exogenous, lactating, exchanges, affects, mammary, gland, ….

Table 2: Sample output of the GoPubMed term extraction process for the Cadherin-related tumor suppressor protein.

[5] We chose such a small sample to facilitate the collection of evidence from GoPubMed, which is not yet fully automated. Our XOA approach is very scalable, and we do not anticipate any problem running the full protein data set of 255,502 pairs, once we fully automate the GoPubMed extraction process.

4 Integrating Textual Evidence in XOA

Using the output of the GoPubMed term extraction process, we created vector-based signatures for each of the 1783 proteins, where

- features are obtained by stemming the terms provided by GoPubMed
- the value for each feature is derived as the *tf*idf* for the feature.

We then calculated the similarity between each of the 2360 protein pairs as the cosine value of the two vector-based signatures associated with the protein pair.

We tried two different strategies to augment the XOA score for protein similarity using the protein similarity values obtained as the cosine of the GoPubMed term-based signatures. The first strategy adopts a fusion approach in which the two similarity measures are first normalized to be commensurable and then combined to provide an interpretable integrated model. A simple normalization is obtained by observing that the Resnik's information content measure is commensurable to

the log of the text based cosine (LC). This leads us to the fusion model shown in (5) for XOA, based on Resnik's semantic similarity measure (XOA_R).

(5) $Fusion(Resnik) = XOA_R + LC$

We then observe that the XOA measures based on Resnik, Lin (XOA_L) and Jiang & Conrath (XOA_{JC}) are highly correlated (correlations exceed 0.95 on the large benchmarking dataset discussed in section 2, see Table 1). This suggests the fusion model shown in (6), where the averages of the XOA scores are computed from the benchmarking data set.

(6) $Fusion(Lin) = XOA_L + LC*Ave(XOA_L)/Ave(XOA_R)$

$Fusion(Jiang\ \&\ Conrath) = XOA_{JC} + LC*Ave(XOA_{JC})/Ave(XOA_R)$

The second strategy consists in building a prediction model for BLAST bit score (BBS) using the XOA score and the log-cosine LC as predictors without the constraint of remaining interpretable. As in the previous strategy, a different model was sought for each of the three XOA variants. In each case, we restrict ourselves to cubic polynomial regression models as such models are quite efficient at capturing complex nonlinear relationships between target and predictors (e.g. Weisberg 2005). More precisely, for each of the semantic similarity measures, we fit the regression model to BBS shown in (7), where the subscript *x* denotes either R, L or JC, and the coefficients *a* to *h* are found by maximizing the Spearman rank order correlations between BBS and the regression model. This maximization is automatically carried out by using a random walk optimization approach (Romeijn 1992). The coefficients used in this study for each semantic similarity measure are shown in Table 3.

(7) $a*XOA_x + b*XOA_x^2 + c*XOA_x + d*LC + e*LC^2 + f*LC^3 + g*XOA_x*LC$

5 Evaluation

Table 4 summarizes the results for both strategies, comparing Spearman rank correlations between BBS and the models from the fusion and regression approaches with Spearman rank correlations between BBS and XOA alone. Note that the latter correlations are lower than the one reported in Table 2 due to the small size of our sample (1% of the original data set, as pointed out above). P-values associated with the changes in the correlation values are also reported, enclosed in parentheses.

	Resnik	**Lin**	**Jiang & Conrath**
a	-10684.43	2.83453e-05	0.2025174
b	1.786986	-31318.0	-1.93974
c	503.3746	45388.66	0.08461453
d	-3.952441	208.5917	4.939535e-06
e	0.0034074	1.55518e-04	0.0033902
f	1.4036e-05	9.972911e-05	-0.000838812
g	713.769	-1.10477e-06	2.461781

Table 3: Coefficients of the regression model maximizing Spearman rank correlation between BBS and the regression model for each of the three semantic similarity measures.

	XOA	**Fusion**	**Regression**
Resnik	0.295	0.325 (>0.20)	0.388 (0.0008)
Lin	0.274	0.301 (>0.20)	0.372 (0.0005)
Jiang & Conrath	0.273	0.285 (>0.20)	0.348 (0.008)

Table 4: Spearman rank order correlation coefficients between BLAST bit score BBS and XOA, BBS and the fusion model, and BBS and the regression model. P-values for the differences between the augmented models and XOA alone are given in parentheses.

An important finding from Table 4 is that integrating text-based evidence in the semantic similarity measures systematically improves the relationships between BLAST and XOA. Not surprisingly, the fusion models yield smaller improvements. However, these improvements in the order of 3% for the Resnik and Lin variants are very encouraging, even though they are not statistically significant. The regression models, on the other hand, provide larger and statistically significant improvements, reinforcing our hypothesis that textual evidence complements the GO-based similarity measures. We expect that a more sophisticated NLP treatment of textual evidence will yield significant improvements even for the more interpretable fusion models.

Conclusions and Further Work

Our early results show that literature evidence provides a significant contribution, even using very simple Information Extraction and integration methods such as those described in this paper. The employment of more sophisticated Information

Extraction tools and integration techniques is therefore likely to bring higher gains.

Further work using GoPubMed involves factoring in the accuracy percentage which related extracted terms to their induced GO categories and capturing complex phrases (e.g. *signal transduction*, *fat protein*). We also intend to compare the advantages provided by the GoPubMed term extraction process with Information Extraction tools created for the biomedical domain such as *Medstract* (Pustejovsky et al. 2002), and develop a methodology for integrating a variety of Information Extraction processes into XOA.

References

Altschul, S.F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Anang, W. Miller and D.J. Lipman (1997) Gapped BLAST and PSI-BLST: a new generation of protein database search programs. *Nucl. Acids Res.* 25:3389-3402.

Andrade, M.A. (1999) Position-specific annotation of protein function based on multiple homologs. *ISMB* 28-33.

Andrade, M.A. and A. Valencia (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *ISMB* 25-32.

Azuaje F., H. Wang and O. Bodenreider (2005) Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies* 2005, pages 9-10.

Bodenreider, O., M. Aubry and A. Burgun (2005) Non-lexical approaches to identifying associative relations in the Gene Ontology. In *Proceedings of Pacific Symposium on Biocomputing*, pages 104-115.

Buckley, C. (1985) Implementation of the SMART information retrieval system. *Technical Report 85-686*, Cornell University.

Budanitsky, A. (1999) Lexical semantic relatedness and its application in natural language processing. Technical report CSRG-390, Department of Computer Science, University of Toronto.

Chang, J.T., S. Raychaudhuri, and R.B. Altman (2001) Including biological literature improves homology search. In *Proc. Pacific Symposium on Biocomputing*, pages 374—383.

Couto, F. M., M. J. Silva and P. Coutinho (2003) Implementation of a functional semantic similarity measure between gene-products. *Technical Report*, Department of Informatics, University of Lisbon, http://www.di.fc.ul.pt/tech-reports/03-29.pdf.

Delfs, R., A. Doms, A. Kozlenkov, and M. Schroeder. (2004) GoPubMed: ontology based literature search applied to Gene Ontology and PubMed. In *Proc. of German Bioinformatics Conference*, Bielefeld, Germany. LNBI Springer.

Doms, A. and M. Schroeder (2005) GoPubMed: Exploring PubMed with the GeneOntology. *Nucleic Acids Research.* 33: W783-W786; doi:10.1093/nar/gki470.

Jiang J. and D. Conrath (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.

Romeijn, E.H. (1992) *Global Optimization by Random Walk Sampling Methods*. Tinbergen Institute Research Series, Volume 32. Thesis Publishers, Amsterdam.

Lin, D. (1998) An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

Lord P.W., R.D. Stevens, A. Brass, and C.A. Goble (2002) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19(10):1275-1283.

Lord P.W., R.D. Stevens, A. Brass, and C.A. Goble (2003) Semantic similarity measures as tools for exploring the Gene Ontology. In *Proceedings of Pacific Symposium on Biocomputing*, pages 601-612.

MacCallum, R. M., L. A. Kelley and Sternberg, M. J. (2000) SAWTED: structure assignment with text description--enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics* **16**, 125-9.

Pearson, W. R. and D. J. Lipman (1988) Improved tools for biological sequence analysis. In *Proceedings of the National Academy of Sciences* 85:2444-2448.

Pedersen, T., S. Banerjee and S. Patwardhan (2005) Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. University of Minnesota Supercomputing Institute Research Report UMSI 2005/25, March. Available at http://www.msi.umn.edu/general/Reports/rptfiles/2005-25.pdf.

Posse, C., A. Sanfilippo, B. Gopalan, R. Riensche, N. Beagley, and B. Baddeley (2006) Cross-Ontological Analytics: Combining associative and hierarchical relations in the Gene Ontologies to assess gene product similarity. To appear in *Proceedings of International*

Workshop on Bioinformatics Research and Applications. Reading, U.K.

Pustejovsky, J., J. Castaño, R. Saurí, A. Rumshisky, J. Zhang, W. Luo (2002) Medstract: Creating large-scale information servers for biomedical libraries. *ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia, PA.

Resnik, P. (1995) Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal.

Sanfilippo A., C. Posse and B. Gopalan **(**2004) Aligning the Gene Ontologies. In *Proceedings of the Standards and Ontologies for Functional Genomics Conference 2*, Philadelphia, PA, http://www.sofg.org/meetings/sofg2004/Sanfilippo.ppt.

Salton, G., A. Wong and C. S. Yang (1975) A Vector space model for automatic indexing, *CACM* 18(11):613-620.

Smith, T. and M. S. Waterman (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.

Weisberg, S. (2005) *Applied linear regression*. Wiley, New York.

Yang, Y. and J.O. Pedersen (1997) A comparative Study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning* (ICML), pages 412-420, Nashville.