A Relevance Detection Approach to Gene Annotation

Wen-Juan Hou, **Chih Lee**, **Kevin Hsin-Yih Lin** and **Hsin-Hsi Chen**
{wjhou, **clee**, **hylin}@nlg.csie.ntu.edu.tw**; **hhchen@csie.ntu.edu.tw**

Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei, Taiwan 106

Abstract

Gene Ontology (GO) enables scientists to describe and annotate gene products with three controlled vocabularies. However, the nature of variation in terminology makes automatic annotation of gene products based on biomedical literature challenging. In this paper, gene annotation was modeled as relevance detection, and an information retrieval with reference corpus was proposed to annotate a gene product with a GO term given a piece of the evidence text. Gene Reference into Functions (GeneRIFs) in NCBI LocusLink database served as the source of evidence in this study. Evidence text, and GO terms along with their definitions were regarded as queries to a reference corpus, which consists of 525,936 MEDLINE abstracts. The similarity between retrieved results measured the degrees of relationship between evidence text and GO terms, and thus guided the annotation. Different number of predicted GO terms, and different distances between predicted and correct terms in GO hierarchy were considered in this study. The results showed that the best recall rate was 78.2% at distance 12 with 5 predicted GO terms, and the best precision rate was 66.2% at distance 12 with one predicted term, when 200 relevant documents were returned by Okapi information retrieval system.

1 INTRODUCTION

As the number of biological and medical publications rapidly grows, the search for desired information becomes more and more difficult. This is further hindered by the wide variations in terminology. Gene Ontology (Ashburner *et al*., 2000) was thus constructed to address the need for standardized descriptions of gene products in different databases. While the development of ontologies is indispensable, scientists cannot benefit from those constructed ontologies until the ontologies are put to broad use. GO is currently adopted by many model organism databases (http://www.geneontology.org/GO. consortiumlist.html) to annotate gene products. But the annotation process requires curators to look into the articles passing some simple filtering processes. Methods for speeding up or automating the annotation process to meet the large volume of literature are thus worthy of investigation.

Since annotating GO terms semi-automatically/ automatically is important, there were some competitions concerning with GO annotations recently. For example, the BioCreative workshop 2004 (http://www.pdg.cnb.uam.es/BioLink/ workshop_BioCreative_04/) initiated a task addressing the assignment of GO annotations to human proteins. It required the participants to automatically annotate a protein with GO terms according to the information found in a publication. The participants also needed to provide the evidence text. The categorization task of the TREC 2004 Genomics Track (http://medir.ohsu.edu/~genomics/) also focused on GO. The annotation subtask is simplified (i.e., not to annotate the precise GO terms) to assigning one or more GO main categories (“biological process”, “cellular component” and “molecular function”) to the articles.

For the automated assignment of GO terms to sequences, the Gene Ontology Annotation (GOA) project (Camon *et al*., 2003) developed mappings between protein domains and GO terms, and between SWISS-PROT (Boeckmann *et al*., 2003) keywords and GO terms. The sequence can be automatically labeled with certain GO terms after it has been annotated with a SWISS-PROT keyword. Poulito *et al*. (2001) and Xie *et al*. (2002) attempted to find function relations between GO terms and genes from the scientific literature. Perez *et al*. (2004) proposed a method to establish the mappings between GO and terms from the MEDLINE database of scientific literature. Some researchers (Ray and Craven, 2004; Verspoor *et al*., 2004) tried to expand the GO terms by finding related words. These approaches slightly improved the recall or coverage rates.

Recently, several tools that made use of GOs were evolving (Al-Shahrour *et al*., 2004; Doniger *et al*., 2003; Draghici *et al*., 2003). It indicated that researchers were interested in exploring the gene with GO terms. The GO consortium also created a link between the known genes and the associated GOs, without providing the evidence text in the literature. In this paper, we provided some GO candidate terms accompanying with the evidence. This could help the annotators speed up their work because they no longer have to read full texts. GeneRIF in the LocusLink database (http://www.ncbi.nlm.nih.gov/LocusLink/) provides a simple mechanism to allow scientists to add the functional annotation of genes. Treating the GeneRIF of a gene product as a piece of supporting evidence, we tried to find the suitable GO terms.

Intuitively, we can compute the similarity between GeneRIF and GO terms based on the number of matching words. However, GeneRIFs and GO terms are too short to reflect complete concepts even with GO definitions included. Here, we introduce the information retrieval technology to deal with this issue. The postulation is: if the document sets retrieved by a GeneRIF and a GO concept are similar, they are considered to be relevant to each other, and thus a link can be established in between. For example, the literature with PMID 11798066 was referenced by a GeneRIF of gene "Dag1", and it was also referenced by three GO terms – say, "protein binding", "morphogenesis of an epithelial sheet" and "dystroglycan complex". Meanwhile, "protein binding" comes from the GO category "molecular function", "morphogenesis of an epithelial sheet" belongs to the category "biological process" and "dystroglycan complex" locates in the category "cellular component". It shows that the postulation of relating GeneRIFs and GO terms with all GO categories may be reasonable.

The rest of this paper is organized as follows. In Section 2, we present the flow of our annotating procedure. The basic idea and the experimental methods in this study are introduced in Section 3. Section 4 shows the results and makes some discussions. Finally, Section 5 contains concluding remarks and suggests the direction of future research.

2 ANNOTATION FLOW

Generally, a gene name may have several aliases. Different functions for a gene may be discovered in different articles. For example, "Gli3" has aliases "Xt", "Bph", "Pdn" and "add". Meanwhile, there are nineteen functions discovered from several documents (e.g., PMID 12435627, 12435361, 12435629, *etc*.) for the gene "Gli3". Users interested in the functions of a given gene can consult the existing resources such as GeneRIF in the LocusLink database, or read the newly published articles that are not yet referenced and annotated by the database curators. Relevance detection will help database curators or ontology annotators maintain existing resources and keep them up to date. An annotation system may consist of two major stages - (1) the extraction of molecular function of a gene from the literature and (2) the annotation of this function with a term in a controlled vocabulary (ontology). In the first stage, we extract the evidence text from the literature that will support GO annotation in the second stage. In this paper, MEDLINE abstracts, GeneRIF and GO served as experimental objects.

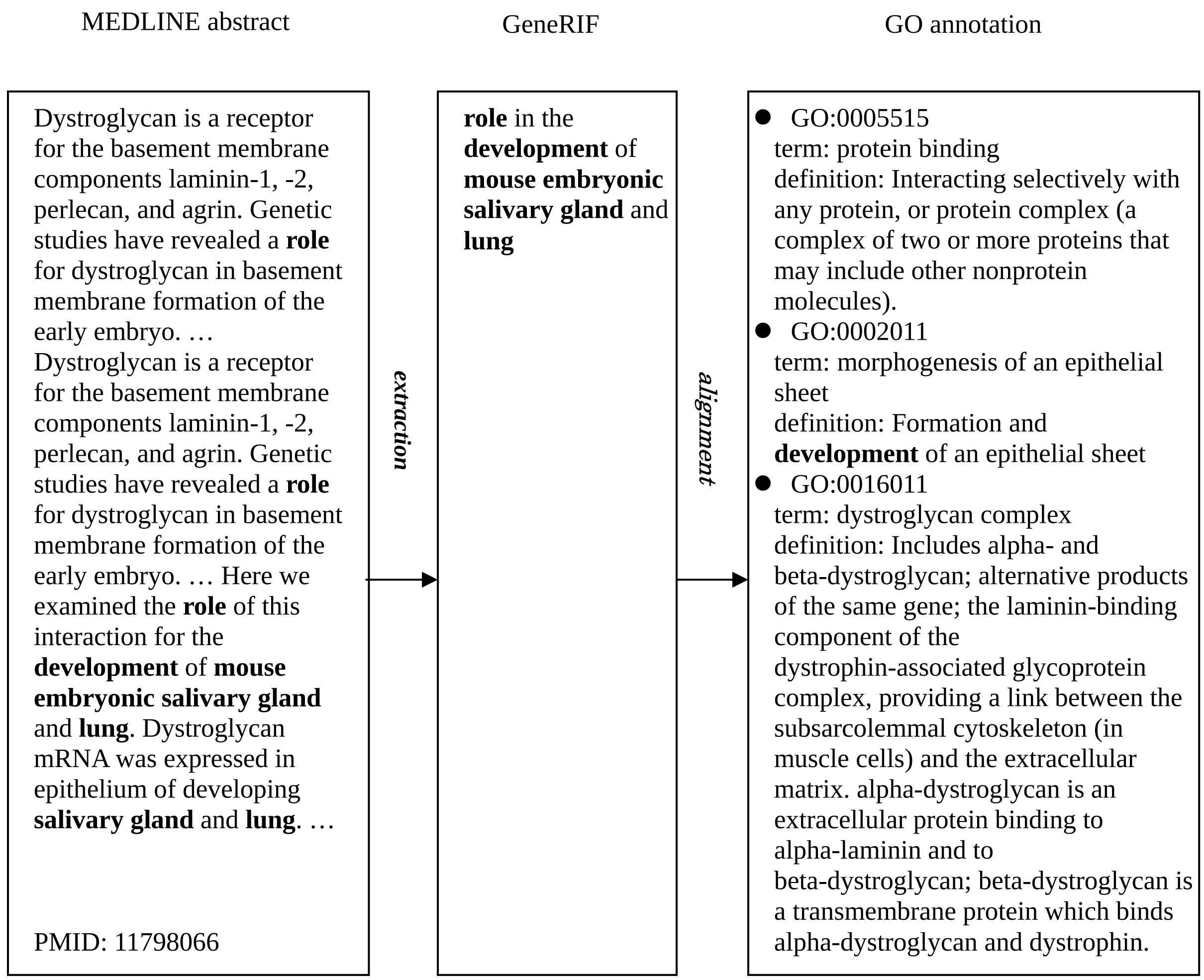
Figure 1 shows an example illustrating our idea. The left part is a MEDLINE abstract with the function description highlighted. The middle part is the corresponding GeneRIF, which is extracted from the partial sentence of the abstract. The right part lists the GO annotations that are annotated by referencing the same abstract. The matching words between MEDLINE and GeneRIF, or between GO and GeneRIF are in bold. The issues in the first stage have been addressed by several researchers (deBruin *et al*, 2003; Jelier *et al*, 2003; Kayaalp *et al*, 2003; Lee *et al*., 2004; Mitchell *et al*, 2003), and are not discussed in this paper. This paper is focused on the second stage. The GeneRIFs in the LocusLink database are used directly in our study. Figure 1 also shows that the number of matching words between GeneRIF and GO concepts is small, so that direct word matching is not practical.

3 METHODS

The outline of the methods is shown in Figure 2. We first prepared the data used in this study. Then, we computed the similarity measures between GeneRIFs and GO terms based on the relevance detection approach. Finally, the predicted GO terms for GeneRIFs were proposed. A detailed description of "Data preparation" is given in Section 3.1. The method for "Similarity computation" with "Reference corpus" will be explained in Section 3.2. Finally, Section 3.3 introduces the methods for generating the predicted GO terms.

3.1 Data Preparation

To find the relation between GO terms and articles, it is straightforward to use the corpus that identifies sentences from the article so that the sentences are the evidence of assigning GO terms. Unfortunately, there exists no such the corpus freely available as we know. Nevertheless, the GeneRIFs in the LocusLink database warrant the assignment of GO terms but it also does not highlight the sentences from the article to tell researchers why the connection is made. Since we regarded the GeneRIFs of a gene product as supporting evidence for the assignment of GO terms, a GeneRIF could only warrant the assignment of GO terms under the assumptions that both the GeneRIFs and the GO terms referenced to the same document.



Dystroglycan is a receptor for the basement membrane components laminin-1, -2, perlecan, and agrin. Genetic studies have revealed a **role** for dystroglycan in basement membrane formation of the early embryo. … Dystroglycan is a receptor for the basement membrane components laminin-1, -2, perlecan, and agrin. Genetic studies have revealed a **role** for dystroglycan in basement membrane formation of the early embryo. … Here we examined the **role** of this interaction for the **development** of **mouse embryonic salivary gland** and **lung**. Dystroglycan mRNA was expressed in epithelium of developing **salivary gland** and **lung**. …

PMID: 11798066

role in the **development** of **mouse embryonic salivary gland** and **lung**

- GO:0005515
  term: protein binding
  definition: Interacting selectively with any protein, or protein complex (a complex of two or more proteins that may include other nonprotein molecules).
- GO:0002011
  term: morphogenesis of an epithelial sheet
  definition: Formation and **development** of an epithelial sheet
- GO:0016011
  term: dystroglycan complex
  definition: Includes alpha- and beta-dystroglycan; alternative products of the same gene; the laminin-binding component of the dystrophin-associated glycoprotein complex, providing a link between the subsarcolemmal cytoskeleton (in muscle cells) and the extracellular matrix. alpha-dystroglycan is an extracellular protein binding to alpha-laminin and to beta-dystroglycan; beta-dystroglycan is a transmembrane protein which binds alpha-dystroglycan and dystrophin.

Fig. 1. An example of complete annotation from the literature to GO.

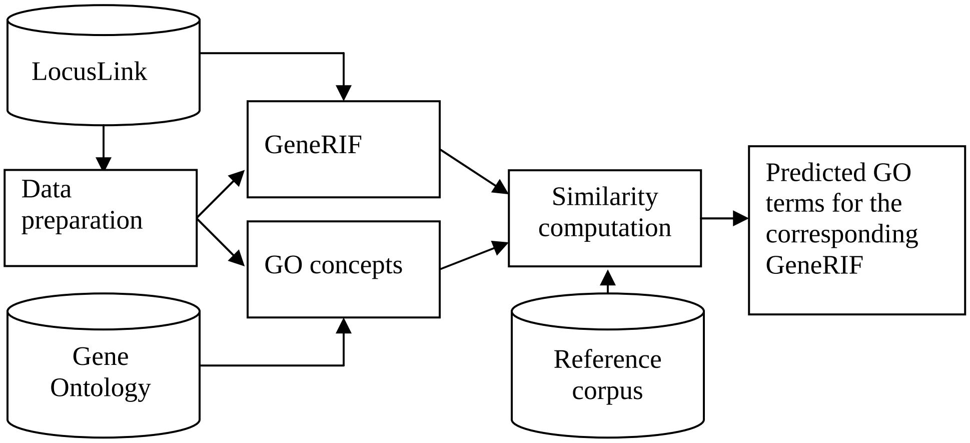


Fig. 2. The flowchart of our methods.

This method of data preparation looked biased but we collected it because GO terms and GeneRIFs indeed existed with some relationship under this assumption. We could find some relationship between them henceforth. The experimental data was therefore collected as follows.

We downloaded the data from the LocusLink database and got 208,877 genes. For each entry in the LocusLink database, if the gene under consideration has both GeneRIFs and GO terms referencing to the same MEDLINE abstracts, the pair of the GeneRIFs and GO terms was collected as supporting evidence and possible answer keys of annotation for this gene, respectively. Furthermore, we examined the corresponding GO terms manually and filtered out those which were obviously unrelated to the corresponding GeneRIFs. The manual checking procedure is to guarantee that the GeneRIFs and GO terms not only come from the same documents, but also have meaningful relationships. Figure 3 shows an example of gene “Dock2”. The symbol * denotes that the corresponding GO terms are filtered out from the possible answer keys. In this way, total 550 pairs of GeneRIFs and GO terms were obtained from 335 distinct genes, and used for testing.

3.2 Similarity Measure

A GeneRIF is often a sentence or a small passage that describes the function of a particular gene product. A GO term is composed of a few words and often comes along with a definition. In the following sections, “a GO term plus its definition” is denoted as “a GO concept”. Intuitively, we can compute the similarity between a GeneRIF and all the GO concepts by keyword matching to set up the linkage. However, the challenge is that GeneRIFs and GO concepts are both too short to provide sufficient information, which makes this approach less promising. In the example shown in Figure 1, only one common stemmed word, i.e., ”development”, appears in both the GeneRIF and one GO concept. To deal with this problem, we introduce the idea of relevance detection in TREC Novelty Track (Harman, 2002). The novelty track aimed at detecting relevance and novelty from a set of sentences, and that required the computation of similarity between two sentences. Tagging molecular functions with GO terms could be modeled as a relevance detection problem. The information retrieval (IR) approach with reference corpus proposed by Chen, Tsai and Hsu (2004) was adopted to resolve this problem in this study.

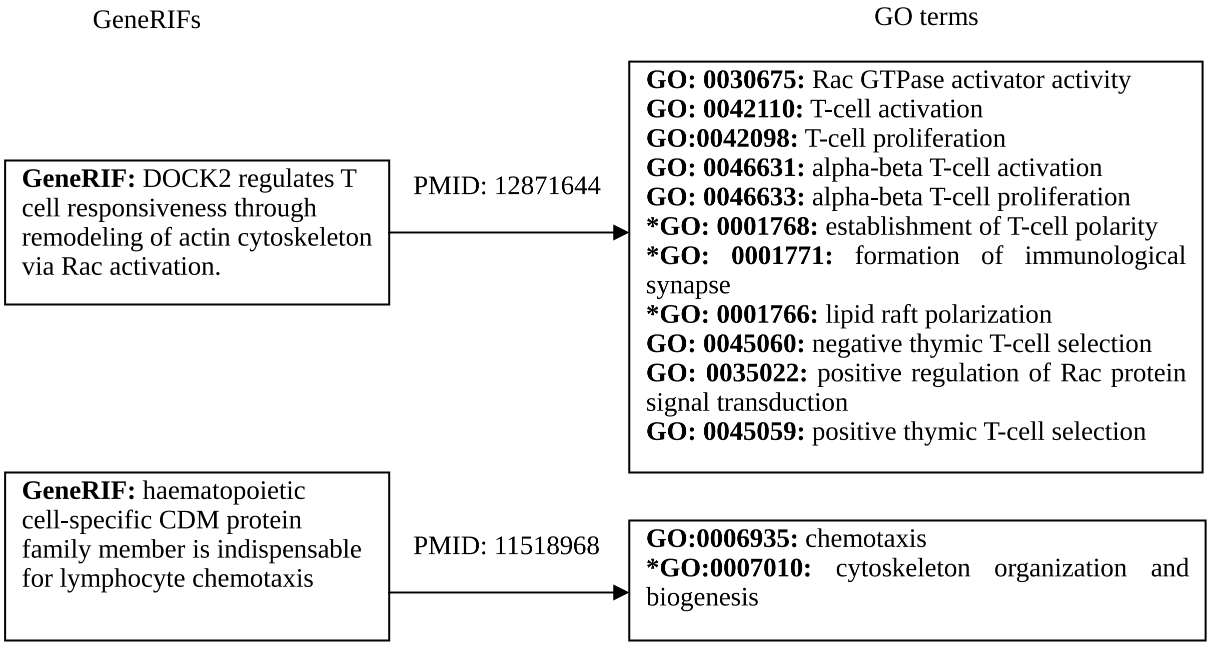

GeneRIFs
GO terms
GeneRIF: DOCK2 regulates T cell responsiveness through remodeling of actin cytoskeleton via Rac activation.
PMID: 12871644
GO: 0030675: Rac GTPase activator activity
GO: 0042110: T-cell activation
GO:0042098: T-cell proliferation
GO: 0046631: alpha-beta T-cell activation
GO: 0046633: alpha-beta T-cell proliferation
*GO: 0001768: establishment of T-cell polarity
*GO: 0001771: formation of immunological synapse
*GO: 0001766: lipid raft polarization
GO: 0045060: negative thymic T-cell selection
GO: 0035022: positive regulation of Rac protein signal transduction
GO: 0045059: positive thymic T-cell selection
GeneRIF: haematopoietic cell-specific CDM protein family member is indispensable for lymphocyte chemotaxis
PMID: 11518968
GO:0006935: chemotaxis
*GO:0007010: cytoskeleton organization and biogenesis

Fig. 3. GeneRIFs and GO terms extracted from gene “Dock2” (LocusID: 94176).

With this IR approach, the similarity/relevance between two sentences is obtained as follows. Each sentence is treated as a query to an IR system, and the top 200 relevant documents along with a weight for each document are retrieved from the reference corpus. The sentence is then represented by a list of weighted documents. Finally, the Cosine function is used to measure the similarity between the two sentences.

$$\cos(v_i, v_j) = \frac{\sum_{k=1}^{n}(v_{i,k} \times v_{j,k})}{\|v_i\| \cdot \|v_j\|}$$

Where v_i and v_j denote two vectors representing the sentences to be compared, and n is the number of documents in the corpus.

Choosing the appropriate reference corpus, the IR system and the strategies for selecting predicted GO terms is important in our experiments. We explored each issue in details as follows.

First, the reference corpus consulted should be large enough to cover different themes for references. In our experiments, the documents used in TREC 2003 Genomics Track were adopted as the reference corpus. The text collection consists of 525,936 MEDLINE records where indexing has been completed between 4/1/2002 and 4/1/2003. Secondly, a highly effective information retrieval system, Okapi (Online Keyword Access to Public Information), was adopted as our platform. In the subsequent experiments, Okapi system with the basic setting of *bm25* (Robertson *et al.*, 1998) was employed. It has the average precision of 0.2253 on the 50 training topics in TREC 2003 Genomics Track (Hersh and Bhupatiraju, 2003). Finally, the methods for selecting the predicted GO terms are investigated in our study and will be discussed in Section 4.

3.3 Generation of Predicted GO Terms

Although GO is organized as a directed acyclic graph, we did not explore and utilize its hierarchical properties at first. The 17,961 GO terms from three main hierarchies were viewed as a flat list of GO terms. Therefore, each GO concept was expanded to a weighted document vector using the aforementioned IR approach. Given a piece of supporting evidence for a gene product, i.e., a GeneRIF, it is first expanded to a weighted document vector using the same IR approach. The similarity between this GeneRIF and every GO concept under the molecular function hierarchy is computed to obtain a ranked list of predicted GO terms with similarity value greater than zero. The process is illustrated in Figure 4.

4 RESULTS AND DISCUSSIONS

GO terms have a hierarchical structure which is appropriate for a flexible annotation process (Ashburner *et al.*, 2000). Perez *et al.* (2004) proposed a metric to evaluate the annotation performance under the GO hierarchy. They analyzed the distance in the GO hierarchy between the GO terms predicted and those in the answer set. The matched distance is *n* if the length of the shortest path between the predicted term and the answer term is *n*. For example, if the predicted term is “minus-end-directed microtubule motor activity” (which depends on “microtubule motor activity” that

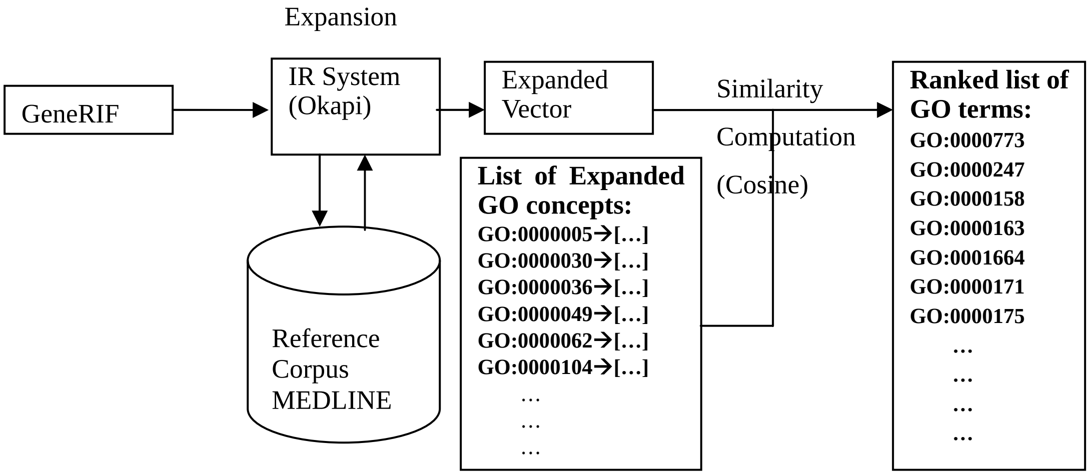


Fig. 4. The flow of generating predicted GO terms.

depends on "motor activity") and the correct term is "microfilament motor activity" (which depends on "motor activity"), then there is a match with a distance of 3 in the hierarchy. We adopted the same evaluation metric as Perez *et al*.'s (2004).

The number of predicted GO terms is also considered in this study. If we propose only one GO term, the curators have only one candidate to choose. However, it is not always the case. Figure 3 shows that GeneRIF (from the document with PMID: 12871644) for "Dock2" can be related to eight GO terms. Thus, the GeneRIF can obviously warrant the assignment of more than one GO term. However, how many candidates we should propose is worth investigating. In the primary experiment, we proposed the predicted GO terms from top 1 to top 5. Figures 5 and 6 show the recall rates and the precision rates under different settings, respectively. Different distances in the GO hierarchy between the predicted terms and the answer terms are evaluated.

The experimental results show the recall rate is better than precision rate if at least two GO terms are proposed. It is reasonable because there is an average of 1.41 GO terms for each GeneRIF in our answer set. It is helpful for human curators because they can collect more possible answers. These two figures demonstrate the same performance trend that "distance" and "recall/precision" are positively correlated, and they reach the stable situation at distance 12. The recall rates show Top5 > Top4 > Top3 > Top2 > Top1 while the precision rates show Top1 > Top2 > Top3 > Top4 > Top5. That is rational because proposing more predicted terms will increase the recall rates but decrease the precision rates. Let's see an example of gene "Slc26a1" with LocusID 231583. "Slc26a1" is annotated with the following four GO terms, i.e., "oxalate transporter activity", "oxalate transport", "chloride transporter activity" and "chloride transport". If we predicted one GO term, "oxalate transporter activity" is found with perfect matching. If two predicted terms were produced, "oxalate transporter activity" and "oxalate transport" were produced with perfect matching. Considering three predicted terms, "anion exchanger activity" was added. It depends on "chloride transporter activity" and therefore, was considered to match at a distance of 1 in the hierarchy. The last one, "chloride transport", was not found in the fourth prediction. However, the fifth prediction was "sulfate transport" that is a child of "inorganic anion transport" and "chloride transport" is a child of "inorganic anion transport". Hence, they are siblings in the process hierarchy. It was the case of a match at distance 2. This example explained why recall rates are increased when the number of predicted terms or the evaluation distance is increased.

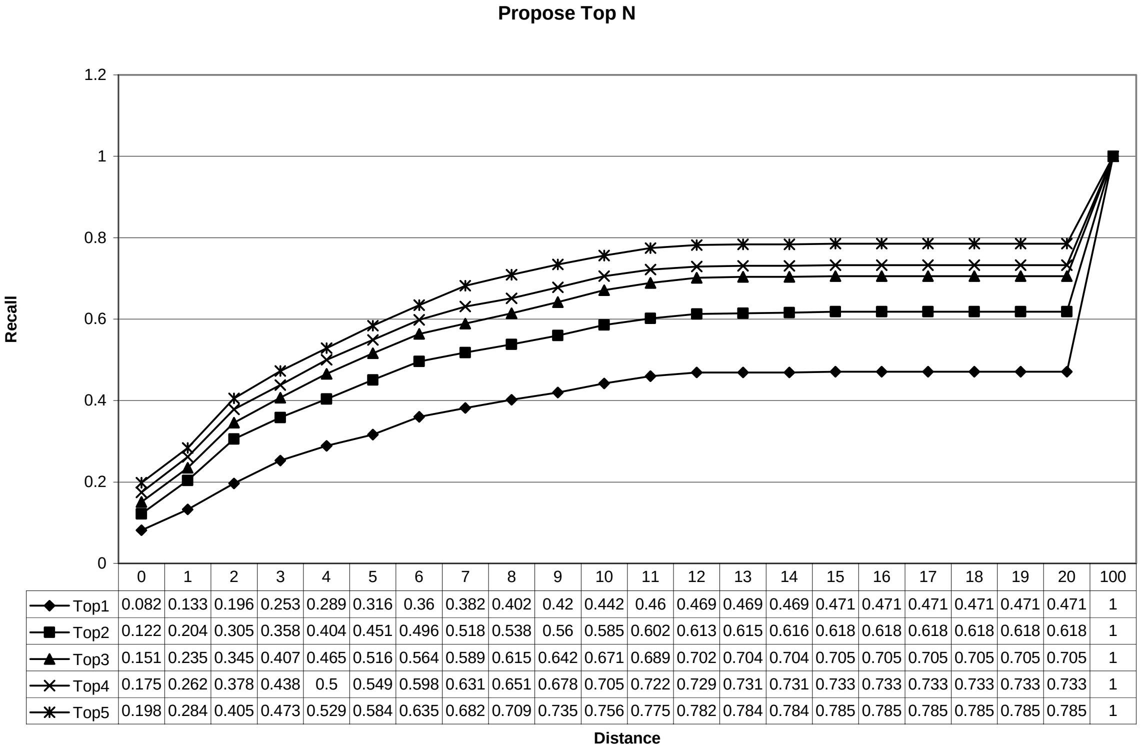
In addition, the best recall rate is 78.2% at distance 12 with five predicted GO terms where the precision rate is 22.1%. The best precision rate is 66.2% at distance 12 with one predicted term, and the corresponding recall rate is 46.9%. If considering the perfect matches and matches at distance 1 in the hierarchy, we got a recall of 28.4% with five predicted terms and a precision of 8.0%. Furthermore, predicting only one term, i.e., perfect matches, resulted in a recall of 13.3% and a precision of 18.7%.

Some related research has been taken before. Let's outline the brief results as follows. Perez *et al.* (2004) annotated GO terms by associating MEDLINE MeSH references with a recall of 8% and a precision of 67%. Pouliot *et al*. (2001) hand-made an ontology on protein domains, and then classified GO terms to their ontology. They did not provide the evaluation results for all GO terms because the evaluation was human-checked. Xie *et al*. (2002) used text information combined with a cellular localization predictive tool for prediction. The evaluation considered only the best predicted GO term and they evaluated by GO categories instead of GO terms. Because the experiments were made on different test sets and the evaluation criteria were different, it had not enough evidence to conclude which methods performed better. Furthermore, since the corpus and the evaluation criteria were different in our work than in previous research, there was no real basis for comparing their results with ours. Here, we made another experiment without reference corpus for comparison, i.e., direct word matching. Considering the example shown in Figure 1, only the word "development" among correct GO concepts was matched under this experiment, and it was possible that words of GeneRIF matched with other incorrect GO concepts. We define the increase ratio as follows to measure the performance difference between methods with/without reference corpus.

$$\text{Increase ratio} = \frac{\text{Prc - Pem}}{\text{Pem}},$$

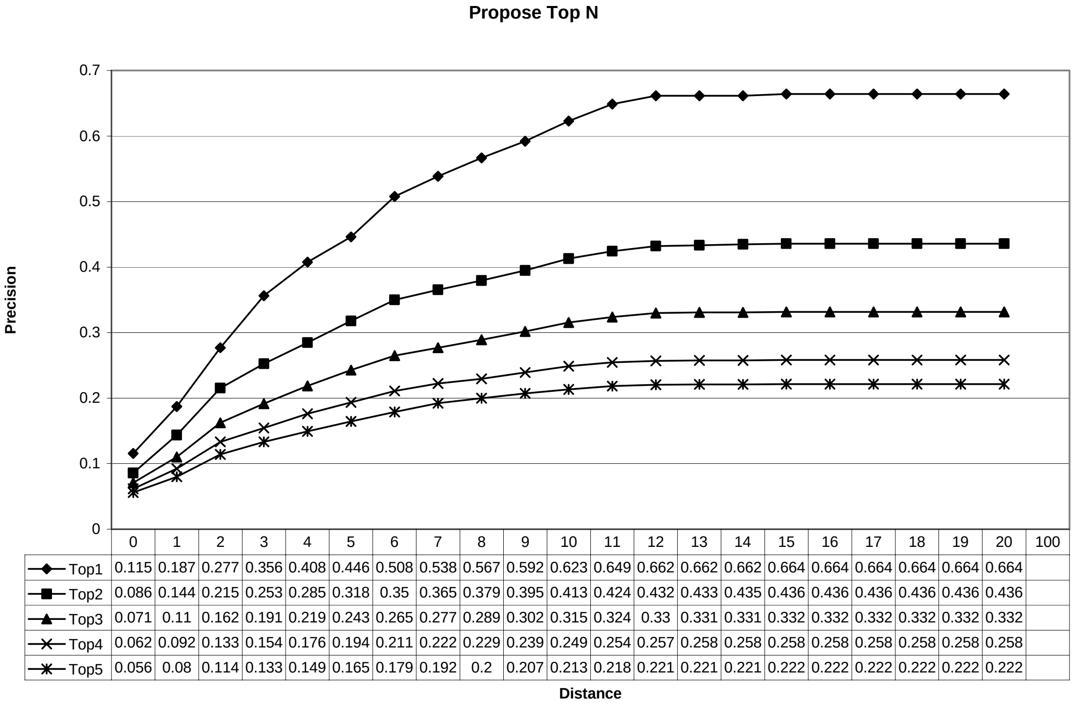
where Prc indicates "Performance with reference corpus" and Pem stands for "Performance with exact matching".

The results of the methods with/without reference corpus are shown in Figures 7 and 8, which illustrated increase ratios of recall rates and precision rates, respectively. We observed that the increase ratios of recall and precision rates are similar. As illustrated in Figures 6 and 7, the increase ratios are between 73.0% and 81.1% at distance 0. It indicates that the reference corpus is quite useful, especially for perfect matches. After distance 8, the performance of these



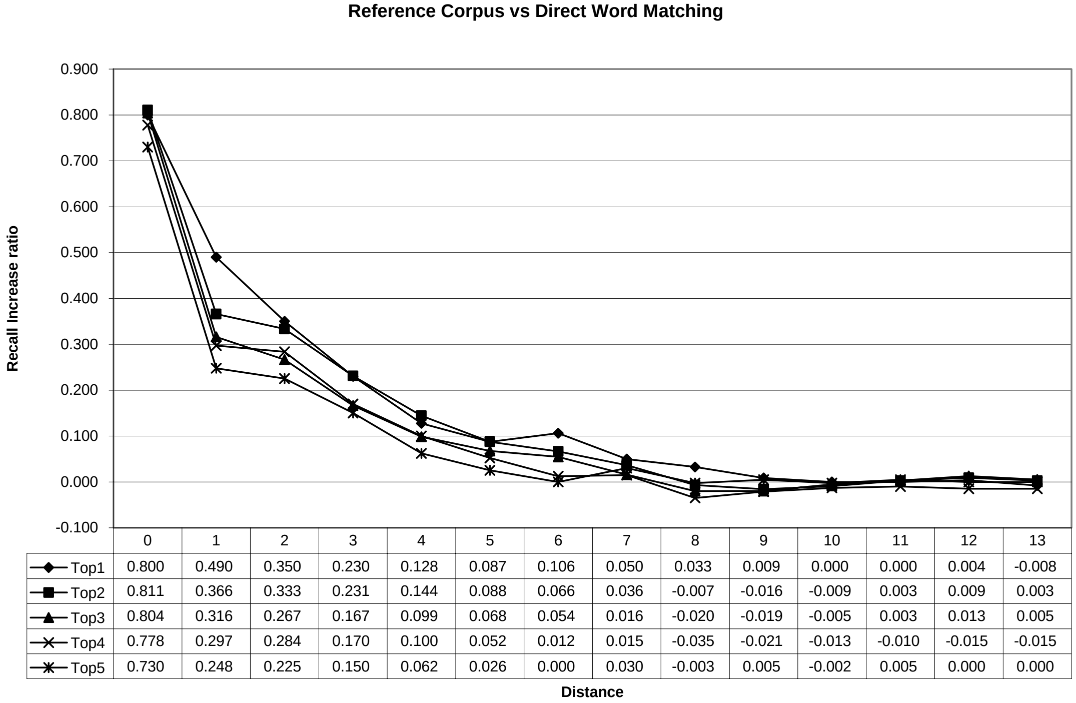
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	100
◆ Top1	0.082	0.133	0.196	0.253	0.289	0.316	0.36	0.382	0.402	0.42	0.442	0.46	0.469	0.469	0.469	0.471	0.471	0.471	0.471	0.471	0.471	1
■ Top2	0.122	0.204	0.305	0.358	0.404	0.451	0.496	0.518	0.538	0.56	0.585	0.602	0.613	0.615	0.616	0.618	0.618	0.618	0.618	0.618	0.618	1
▲ Top3	0.151	0.235	0.345	0.407	0.465	0.516	0.564	0.589	0.615	0.642	0.671	0.689	0.702	0.704	0.704	0.705	0.705	0.705	0.705	0.705	0.705	1
× Top4	0.175	0.262	0.378	0.438	0.5	0.549	0.598	0.631	0.651	0.678	0.705	0.722	0.729	0.731	0.731	0.733	0.733	0.733	0.733	0.733	0.733	1
✳ Top5	0.198	0.284	0.405	0.473	0.529	0.584	0.635	0.682	0.709	0.735	0.756	0.775	0.782	0.784	0.784	0.785	0.785	0.785	0.785	0.785	0.785	1

Fig. 5. Recall rates in different numbers of predicted GO terms.



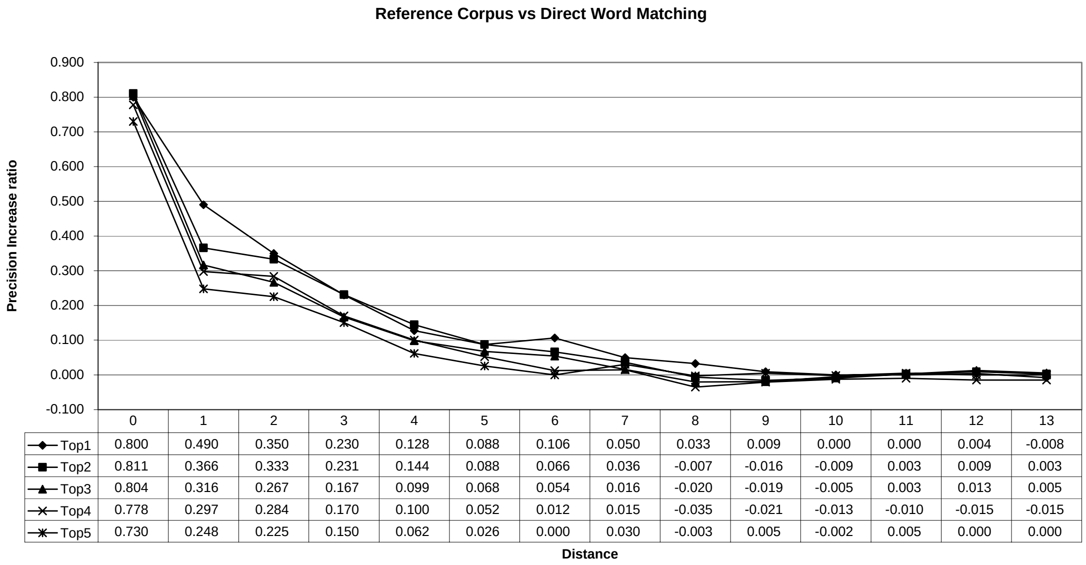
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	100
◆ Top1	0.115	0.187	0.277	0.356	0.408	0.446	0.508	0.538	0.567	0.592	0.623	0.649	0.662	0.662	0.662	0.664	0.664	0.664	0.664	0.664	0.664	
■ Top2	0.086	0.144	0.215	0.253	0.285	0.318	0.35	0.365	0.379	0.395	0.413	0.424	0.432	0.433	0.435	0.436	0.436	0.436	0.436	0.436	0.436	
▲ Top3	0.071	0.11	0.162	0.191	0.219	0.243	0.265	0.277	0.289	0.302	0.315	0.324	0.33	0.331	0.331	0.332	0.332	0.332	0.332	0.332	0.332	
× Top4	0.062	0.092	0.133	0.154	0.176	0.194	0.211	0.222	0.229	0.239	0.249	0.254	0.257	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	
✳ Top5	0.056	0.08	0.114	0.133	0.149	0.165	0.179	0.192	0.2	0.207	0.213	0.218	0.221	0.221	0.221	0.222	0.222	0.222	0.222	0.222	0.222	

Fig. 6. Precision rates in different numbers of predicted GO terms.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
◆ Top1	0.800	0.490	0.350	0.230	0.128	0.087	0.106	0.050	0.033	0.009	0.000	0.000	0.004	-0.008
■ Top2	0.811	0.366	0.333	0.231	0.144	0.088	0.066	0.036	-0.007	-0.016	-0.009	0.003	0.009	0.003
▲ Top3	0.804	0.316	0.267	0.167	0.099	0.068	0.054	0.016	-0.020	-0.019	-0.005	0.003	0.013	0.005
× Top4	0.778	0.297	0.284	0.170	0.100	0.052	0.012	0.015	-0.035	-0.021	-0.013	-0.010	-0.015	-0.015
✳ Top5	0.730	0.248	0.225	0.150	0.062	0.026	0.000	0.030	-0.003	0.005	-0.002	0.005	0.000	0.000

Fig. 7. Increase ratios of recall rates with/without reference corpus at different distances.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
◆ Top1	0.800	0.490	0.350	0.230	0.128	0.088	0.106	0.050	0.033	0.009	0.000	0.000	0.004	-0.008
■ Top2	0.811	0.366	0.333	0.231	0.144	0.088	0.066	0.036	-0.007	-0.016	-0.009	0.003	0.009	0.003
▲ Top3	0.804	0.316	0.267	0.167	0.099	0.068	0.054	0.016	-0.020	-0.019	-0.005	0.003	0.013	0.005
× Top4	0.778	0.297	0.284	0.170	0.100	0.052	0.012	0.015	-0.035	-0.021	-0.013	-0.010	-0.015	-0.015
✳ Top5	0.730	0.248	0.225	0.150	0.062	0.026	0.000	0.030	-0.003	0.005	-0.002	0.005	0.000	0.000

Fig. 8. Increase ratios of precision rates with/without reference corpus at different distances.

two approaches is nearly of no difference. It shows that the information retrieval approach with reference corpus has better performance within shorter distances. In the statistics of GO branching, there are maximum 522 branches and minimum one branch. The average is 4.17 with standard deviation 11.33.

Furthermore, the most frequent depth is 9. In such a case, GO curators will not want to examine the predicted terms with large distances because long distances of the GO hierarchy means that the trace scope to determine the correct GO terms by curators is also wide.

5 CONCLUDING REMARKS

This paper models gene annotation as a relevance detection problem. We proposed a semi-automatic way to assign a GO term to a gene based on its evidence description. Instead of explicitly expanding the GO terms, we relied on an IR system and a reference corpus to expand both GeneRIFs and GO terms implicitly. In the proposed IR approach with reference corpus, a GeneRIF and a GO concept are regarded as two sentences. Similarity between the two sentences determines the assignment of a GO term to the gene product under consideration. This idea, borrowed from the relevance detection in TREC novelty track, is the first attempt to ontology annotation.

The preliminary experiments showed that the promising results will be helpful for the annotation task because only a limited amount of predicted GO terms rather than a complete set of 17,961 terms were proposed for further selection by curators. The improvement space is still open. Machine learning approaches for mining the relationship between GeneRIFs and GO concepts may be explored. Furthermore, combining the other approaches with ours may increase the performance and that will be the future work.

ACKNOWLEDGEMENTS

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC 93-2213-E-002-078 and NSC 93-2752-E-001-001-PAE.

REFERENCES

Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004) FatiGO: a Web Tool for Finding Significant Associations of Gene Ontology Terms with Groups of Genes, *Bioinformatics*, **20**, 578-580.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M. *et al.* (2000) Gene Ontology: Tool for the Unification of Biology, *Nature Genetics*, **25**, 25-29.

BioCreative Workshop. http://www.pdg.cnb.uam.es/BioLink/workshop_BioCreative_04/

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT Protein Knowledgebase and its Supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365-370.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A. and Apweiler, R. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research*, **13**, 1-11.

Chen, H.H., Tsai, M.F. and Hsu, M.H. (2004) Identification of Relevant and Novel Sentences, *Proceedings of 26th European Conference on Information Retrieval*, Lecture Notes in Computer Science, **2997**, Springer-Verlag, 85-98.

deBruin, B. and Martin, J. (2003) Finding Gene Function Using LitMiner. *The Twelfth Text Retrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Doniger, S., Salomonis, N., Dahlquist, K., Vranizan, K., Lawlor, S. and Conklin, B. (2003) MAPPFinder: Using Gene Ontology and GenMAPP to Create a Global Gene-expression Profile from Microarray Data, *Genome Biology*, **4**(1): R7.

Draghici, S., Khatri, P., Bhavsar P., Shah, A., Krawetz, S. and Tainsky, M. (2003) Onto-Tools, the Toolkit of the Modern Biologist: Onto-Express, Onto-Compare, Onto-Design and onto-Translate. *Nucleic Acids Research*, **31**, 3775-3781.

Gene Ontology Consortium. http://www.geneontology.org/GO.consortiumlist.html

Harman, D. (2002) Overview of the TREC 2002 Novelty Track, *Proceedings of TREC 2002*.

Hersh, W. and Bhupatiraju, R.T. (2003) TREC Genomics Track Overview, *Proceedings of TREC 2003*.

Jelier, R., Schuemie, M., *et al*. (2003) Searching for GeneRIFs: Concept-based Query Expansion and Bayes Classification. *The Twelfth Text Retrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Kayaap, M., Aronson, A., *et al*. (2003) Methods for Accurate Retrieval of MEDLINE Citations in Functional Genomics. *The Twelfth Text Retrieval Conference: TREC 2003*, Gaithersburg, MD. NIST.

Lee, C., Hou, W.J. and Chen, H.H. (2004) Support Vector Machine Approach to Extracting Gene References into Function from Biological Documents, *Proceedings of COLING 2004: Joint Workshop on Natural Language Processing in Biomedicine and its Applications*.

LocusLink database. http://www.ncbi.nlm.nih.gov /LocusLink/

Mitchell, J., Aronson, A., *et al*. (2003) Gene Indexing: Characterization and Analysis of NLM's GeneRIFs. *Proceedings of the AMIA 2003 Annual Symposium*, Washiongton, DC. Hanley & Belfus, 460-464.

Perez, A.J., Perez-Iratxeta, C., Bork, P., Thode, G. and Andrade, M.A. (2004) Gene Annotation from Scientific Literature Using Mappings between Keyword Systems. *Bioinformatics*, **20(13)**, 2084-2091.

Pouliot, Y., Gao, J., Su, Q.J., Liu, G.G. and Ling, X.B. (2001) DIAN: a Novel Algorithm for Genome Ontological Classification. *Genome Research*, **11** 1766-1779.

Ray, S. and Craven, M. (2004) Learning Statistical Models for Annotating Proteins with Function Information using Biomedical Text, *BioCreative Workshop Handouts*.

Robertson, S.E., Walker, S. and Beaulieu, M. (1998) Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive, *Proceedings of the Seventh Text Retrieval Conference*, 253-264.

TREC Genomics Track. http://medir.ohsu.edu /~genomics/

Verspoor, K., Cohn, J., Joslyn, C., Mniszewski, S., Rechtsteiner, A., Rocha, L.M., and Simas, T. (2004) Protein Annotation as Term Categorization in the Gene Ontology, *BioCreative Workshop Handouts*.

Xie, H., Wasserman, A., Levine, Z., Novik, A., Grebinskiy, V., Shoshan, A. and Mintz, L. (2002) Large-scale Protein Annotation through Gene ontology. *Genome Research*, **12** 785-794.