Gene Ontology Annotation Using Word Proximity Relationship

Kevin Hsin-Yih Lin, Wen-Juan Hou and Hsin-Hsi Chen
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
{hylin, wjhou}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

In this paper, we propose an approach for doing Gene Ontology (GO) annotation on full-text biomedical articles. This system explores the word proximity relationship between genes and GO terms. We associate genes and GO terms by considering the density function between gene-GO pairs in a paragraph. Different density models are built and several evaluation criteria are employed to assess the effects of the proposed methods. In the best case, we got a precision of < 88% and a recall of < 12%.

1 Introduction

A large amount of biological and medical data is stored in various databases. How to integrate the information of interesting genes scattered around different databases is an important research issue. Gene Ontology (GO) (Ashburner *et al*., 2000; The Gene Ontology Consortium, 2001; Ashburner and Lewis, 2002) is one of the databases that focus on providing standard vocabularies of gene products in different databases. In GO, there are three kinds of structured controlled vocabularies (sub-ontologies) to describe three semantic types of concepts, including molecular function, biological process and cellular component. The sub-ontology represents different categories of genomic characteristics which are described by GO terms. Each GO term is associated with a "GO ID". For example, a GO term "biological_process" has the "GO ID" of "GO:0008150". Several databases (e.g., SGD[1], Flybase[2] and MGI[3]) that belong to different model organism databases annotate their gene products with GO terms, and provide references as well as indicate what kind of evidences is available to support the annotations. But the annotation process requires curators to look through articles. Methods for speeding up or automating the annotation process to meet the large volume of literature are thus worthy of investigation.

Because of the importance of automatic annotation of GO terms, there were some competitions on GO annotations recently. For example, TREC 2004 and TREC 2005 Genomics Track organized GO categorization tasks.[4] The former simplified GO annotation (i.e., not to annotate the precise GO terms) to the task of assigning one or more GO main categories ("biological process", "cellular component" and "molecular function") to articles, while the latter included three more triage topics. The increase in the number of participants at Genomics Track shows that GO annotation problems attracted a lot of attention.

Several attempts have focused on GO annotation. 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. Joslyn *et al*. (2004) developed the Gene Ontology Categorizer (GOC) to summarize or categorize a list of genes of interest. Their evaluation criteria were different from precision/recall measures. Perez *et al*. (2004) proposed a method for establishing mappings between GO and terms from the MEDLINE database of scientific literature, with a recall of 8% and a precision of 67%. Hou *et al*. (2005) modeled GO annotation as relevance detection and showed 78% recall rates and 66% precision rates at distance 12. Some researchers (Ray and Craven, 2005; Verspoor *et al*., 2005) tried to expand the GO terms by finding related words. These approaches slightly improved recall or coverage rates.

---

[1] http://www.yeastgenome.org/
[2] http://flybase.bio.indiana.edu/
[3] http://www.informatics.jax.org/

[4] http://ir.ohsu.edu/genomics

Most of the previous works used information extracted from the title, abstract and MeSH terms only. Obviously, full-text articles contain more information than abstracts, but they also introduce more noises. This is a challenge we must face when doing GO annotation on full-text articles. One important feature we can extract is the word proximity relationship between genes and GO terms. The postulation is: if only one gene and GO terms appear in the same paragraph, they are considered to be associated to each other. If more than one gene is found, the gene with closer proximity to GO terms is preferred. Consider the GO term "cytoplasm" (GO:0005737) in the literature with PMID 10037727 as follows:

> … There was a strong cross-reaction of the anti-<GENE:trr2>Trr2</GENE> antibody reacted with a 36-kDa protein in the total cell homogenates and cytosolic fractions of both strains that is probably due to the presence of <GO:0005737>**cytoplasmic**</GO> <GENE:trr1>Trr1</GENE>, which is 84% identical to <GENE:trr2>Trr2 </GENE>.

There are two genes and one GO with three gene-GO_term occurrences appearing in the above paragraph. The nearest gene to GO term "cytoplasm" is "Trr1" which is annotated with "cytoplasm" in SGD (Saccharomyces Genome Database) (Ball *et al*., 2000) while "Trr2" is farther, and "Trr2" is not annotated with "cytoplasm". It shows that the postulation of associating GO terms with the nearest gene may be reasonable.

To describe the relationship between PMID, GO terms and genes, we use a 3-tuple representation. For example, the yeast "15S_RRNA" in SGD is annotated with GO terms "structural constituent of ribosome" (GO:0005763), "protein biosynthesis" (GO:0003735), "ribosome assembly" (GO:0006412) and "mitochondrial small ribosomal subunit" (GO:0042255). The annotations are referenced to documents with PMID 6261980, 6280192, 2167435 and 6261980, respectively. In our study, we represent this curation as a triple of <PMID, GO ID, GENE> where "PMID" represents PubMed identifier, "GO ID" represents GO category ID, and "GENE" represents the gene name. In the above example, we get four 3-tuples for the yeast "15S_RRNA", i.e., <6261980, 0005763, 15S_RRNA>, <6280192, 0003735, 15S_RRNA>, <2167435, 0006412,

15S_RRNA>, and <6261980, 0042255, 15S_RRNA>.

The rest of this paper is organized as follows. Section 2 sketches an overview of the system architecture. Section 3 depicts how the experimental corpus is built. The details of the proposed methods are presented in Section 4. The experimental results are shown and discussed in Section 5. We also introduce the evaluation metrics in this section. Finally, we make conclusions and present some further work.

2   System Overview

Figure 1 shows the overall architecture of the proposed system. First, we preprocess each full-text article in the corpus, which involves (1) gene name recognition for tagging the gene names, (2) stop-word removal for filtering the stop-words, (3) morphological normalization for getting the stems of verbs and nouns, and (4) GO terms tagging for adding GO ID tags to GO terms. The order of the preprocessing is reasonable. BioTagger is applied first because some gene names contain stop-words and stemming may influence the recognition of gene names. GO term tagging is applied last because we also apply stop-word removal and a stemmer to GO terms. We make use of some biomedical domain specific resources (i.e., BioTagger [5] and Gene Ontology) and some natural language processing resources (i.e., a stop-word list and Porter's stemmer [6]) to preprocess the corpus. After that, we get a set of articles with tagged gene name and tagged GO terms. Then, an algorithm based on word proximity relationship annotates genes with GO terms. Finally, a 3-tuple of <PMID, GO ID, GENE>, which specifies a gene GENE is annotated with some GO term in a biomedical article PMID, is reported.

3   Corpus Construction

To construct an evaluation corpus, we first downloaded all the GO annotation files from the GO website.[7] For each entry in the annotation files, we searched for its corresponding biomedical article using Entrez PubMed.[8] We downloaded the free online full-text articles if

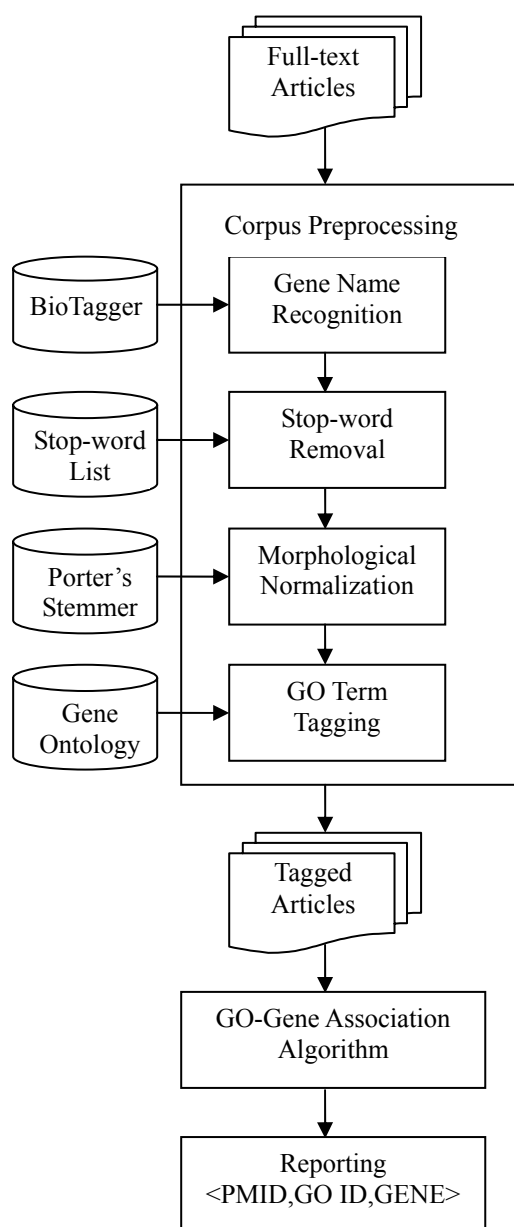they exist. We were able to retrieve 10,054 full-text articles.

Figure 1. System Architecture

Because of the complications with gene name synonyms, we did not use all 10,054 articles. It is common for a gene to have multiple names, so a gene's name in the GO annotation file may be different from its name in a biomedical article. As the focus of this study is not to recognize all the different names of a gene, we decided to filter out the articles which do not contain the gene names as specified in the GO annotation file. This is done by examining each article and keeping the article if at least one entry in the GO annotation file referring to the article also

refers to a gene that appears in the article. After filtering is done, we were left with 4,479 articles. We also removed entries in the GO annotation files which either do not refer to one of these 4,479 articles, or do not refer to a gene name that appears in one of these 4,479 articles. In summary, our final corpus consists of 4,479 full-text biomedical articles, which contain a total of 15,566 annotations.

4 Methods

Our annotation procedure for each article consists of (1) gene name tagging, (2) GO term tagging, and (3) GO term to gene name association. To illustrate our annotation procedure, we give an example of a paragraph in the article with PMID 10198058. The example is shown in Figure 2.

4.1 Gene Name Tagging

We used BioTagger (Liu *et al.*, 2004) to identify all appearances of gene names in an article. BioTagger is a biological entity tagging system capable of recognizing gene names, genomic variations in cancers and malignancy types in cancers. BioTagger has a precision of 77% and a recall of 96% for the yeast. Figure 3 shows the tagging result of the paragraph in Figure 2.

4.2 GO Term Tagging

We used word matching to identify GO terms. In the beginning, we used PubMed's stop-word list to remove all the stop-words from every GO term.[9] We then stemmed the GO terms with Porter's stemmer (Porter, 1980). At the end of these steps, we ended up with a list of processed GO terms which contain stemmed words and no stop-word.

When tagging GO terms, we treated each paragraph of an article as an independent unit. For each paragraph, we went through the list of the processed GO terms to check whether the paragraph contains all the words (named GO-component in this paper) of any particular GO term. If the paragraph did, we considered the paragraph to contain an instance of that particular GO term. The GO-components did not have to appear next to each other or in any particular order in the paragraph. We labeled all the appearances of GO-components in the paragraph with the GO term's GO ID. In brief,

---

9

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.table.pubmedhelp.T38

The Bud Neck Localization of Yck2 Begins as a Ring at the End of Mitosis and Becomes a Patch under the Septum As mentioned previously, bud and mother cell membranes of all large-budded cells with two DAPI-staining regions are equally fluorescent, but a bright ring or thin bright bar is often visible at the bud neck (Figure 2, F-H). The observation of a septum by Nomarski optics that lies above the bright bar (as shown for a diploid cell in Figure 6 B, bottom panel) supports the idea that GFP-Yck2p becomes enriched in the membrane that underlies the growing septum. This change in Yck2p distribution could also parallel the timing of a major change in actin cytoskeletal organization (Adams and Pringle, 1984; Kilmartin and Adams, 1984; Botstein et al., 1997). Before cytokinesis, actin becomes concentrated in a contractile ring at the bud neck (Epp and Chant, 1997; Bi et al., 1998; Lippincott and Li, 1998). As cytokinesis and secretion of new cell wall material to the neck region occur, the cortical actin becomes distributed in patches underlying mother and bud sides of the division site.

Figure 2. Untagged Paragraph with PMID 10198058

The Bud Neck Localization of <GENE>Yck2</GENE> Begins as a Ring at the End of Mitosis and Becomes a Patch under the Septum As mentioned previously, bud and mother cell membranes of all large-budded cells with two DAPI-staining regions are equally fluorescent, but a bright ring or thin bright bar is often visible at the bud neck (Figure 2, F-H). The observation of a septum by Nomarski optics that lies above the bright bar (as shown for a diploid cell in Figure 6 B, bottom panel) supports the idea that <GENE >GFP-Yck2p</GENE> becomes enriched in the membrane that underlies the growing septum. This change in Yck2p distribution could also parallel the timing of a major change in <GENE>actin</GENE> cytoskeletal organization (Adams and Pringle, 1984; Kilmartin and Adams, 1984; Botstein et al., 1997). Before cytokinesis, <GENE>actin</GENE> becomes concentrated in a contractile ring at the bud neck (Epp and Chant, 1997; Bi et al., 1998; Lippincott and Li, 1998). As cytokinesis and secretion of new cell wall material to the neck region occur, the cortical actin becomes distributed in patches underlying mother and bud sides of the division site.

Figure 3. Example of Gene Name Tagging

The <GO:0005935|GO:0007114>**Bud**</GO> <GO:0005935>**Neck**</GO> Localization of <GENE>Yck2</GENE> Begins as a Ring at the End of Mitosis and Becomes a Patch under the Septum As mentioned previously, <GO:0005935|GO:0007114>**bud**</GO> and mother <GO:0007114>cell</GO> membranes of all large-budded <GO:0007114>cells</GO> with two DAPI-staining regions are equally fluorescent, but a bright ring or thin bright bar is often visible at the <GO:0005935|GO:0007114>**bud**</GO> <GO:0005935>**neck**</GO> (Figure 2 , F-H). The observation of a septum by Nomarski optics that lies above the bright bar (as shown for a diploid <GO:0007114>cell</GO> in Figure 6 B , bottom panel) supports the idea that <GENE >GFP-Yck2p</GENE> becomes enriched in the membrane that underlies the growing septum. This change in Yck2p distribution could also parallel the timing of a major change in <GENE>actin</GENE> cytoskeletal organization (Adams and Pringle, 1984; Kilmartin and Adams, 1984; Botstein et al., 1997). Before <GENE>actin</GENE> cytokinesis, actin becomes concentrated in a contractile ring at the <GO:0005935|GO:0007114>**bud**</GO> <GO:0005935>**neck**</GO> (Epp and Chant, 1997; Bi et al., 1998 ; Lippincott and Li, 1998). As cytokinesis and secretion of new <GO:0007114>cell</GO> wall material to the <GO:0005935>**neck**</GO> region occur, the cortical actin becomes distributed in patches underlying mother and <GO:0005935|GO:0007114>**bud**</GO> sides of the division site.

GO:0005935 : bud neck	GO:0007114 : cell budding

Figure 4. Example of GO Term Tagging*

*The upper part is the result of GO term tagging. The lower part lists the GO terms. A word may be tagged with more than one GO ID when it is a GO-component of more than one GO term. For example, bud.

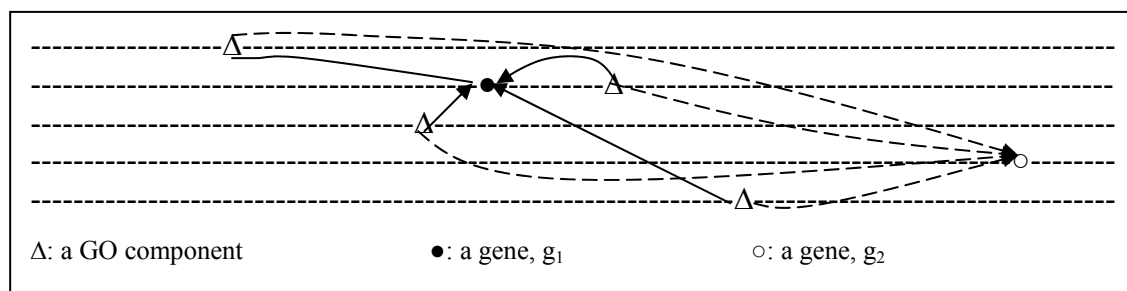Δ: a GO component   ●: a gene, g_1 ○: a gene, g_2

Figure 5. Word Proximity Relationship between Genes and GO-components

a GO-component is any word in a GO term. GO-components are annotated in the text. Since GO-components can be part of different GO terms, each annotation in the text can refer to different GO terms. Figure 4 shows the tagging result of two distinct GO terms. The matching words with GO term "bud neck" (GO:0005935) are in boldface.

4.3 GO to Gene Association

Figure 5 sketches the basic idea of our GO-to-Gene association algorithm. In Figure 5, there are two genes, g_1 and g_2, and four GO-components that may be the same or different in a paragraph. Without loss of generality, we assume there is only one GO term in this paragraph. There is a link between each gene and each GO-component, and it represents the word proximity relationship. We calculate the association scores between them. On the one hand, as we mentioned before, the gene with the closest proximity to the GO term should be associated to the GO term. In other words, if the distance between a gene and a GO-component is shorter, the score is higher. On the other hand, if a GO-component is more important, the score is higher. This model is like a density model: the gene with the highest density (i.e., most tightly surrounded by GO-components) will be selected. To explore the effects between distance and the GO-component's importance, we designed two experiments, Density Models 1 and 2, which are explained in Section 4.3.1 and 4.3.2, respectively.

4.3.1 Density Model 1

For each unique GO term appearing in a paragraph, we associated exactly one gene with it. First, we explore the effect of distance between GO-components and genes. The GO-to-Gene association algorithm is stated informally as follows.

> For each occurrence of a gene in the paragraph, we compute the distance between the gene occurrence and GO-component. The shorter the distance is, the higher the score is. After that, we average the scores of the gene and each GO-component. Then we average the scores for all the gene's occurrences. The gene with the highest score is associated with the GO term.

To describe the above algorithm more formally, we define the following symbols.

G_i: the i-th gene occurrence in a given paragraph,
T_j: the GO term with the j-th unique GO ID,
$T_{j,k}$: the k-th occurrence of a GO term T_j's GO-component in a given paragraph,
$w_{G_i,T_{j,k}}$: total number of words between gene G_i and GO-component $T_{j,k}$, and
c_j: total occurrences of T_j's GO-components in a given paragraph.

Consider a paragraph with n genes in an order of G_1, G_2, …, G_n, where G_i and G_j may be the same or different gene names. Then, the score between G_i and T_j is $s_{G_i,T_j} = \frac{1}{c_j} \sum_{k=1}^{c_j} \frac{1}{w_{G_i,T_{j,k}}}$.

For a certain GO term T_j, the average score of all genes' occurrences identifying the same gene G_i is $avg(s_{G_i,T_j}) = \frac{1}{m} \sum_{l=1}^{m} s_{G_{i_l},T_j}$, for all $G_{i_l} = G_i$ and there are m occurrences with the same name G_i. Finally, a gene G_p with the highest score of $avg(s_{G_p,T_j})$ will be associated with T_j. In other words, G_p and T_j make the most preferred association.

We apply our algorithm to the example in Figure 4. Note that the target GO term T_1 is "bud neck" where the GO-components "bud" and "neck" are in boldface. First of all, there is a single occurrence of gene "Yck2", a single occurrence of gene "GFP-Yck2p" and two occurrences of gene "actin". G_1, "Yck2", has the distances of 3, 2, 19, 48, 49, 136, 137, 161, 174, to the nine occurrences of T_1's GO-components. The association score between G_1 and T_1 is:

$$s_{G_1,T_1} = \frac{1}{9}(\frac{1}{3} + \frac{1}{2} + \frac{1}{19} + \frac{1}{48} + \frac{1}{49} + \frac{1}{136} + \frac{1}{137} + \frac{1}{161} + \frac{1}{174}).$$

Since there is only one occurrence for gene "Yck2", we obtain $avg(s_{G_1,T_1}) = s_{G_1,T_1}$. We compute s_{G_2,T_1}, s_{G_3,T_1} and s_{G_4,T_1} in a similar way where G_2="GFP-Yck2p" and $G_3=G_4$="actin". Moreover, the average scores of each gene are $avg(s_{G_2,T_1}) = s_{G_2,T_1}$, and $avg(s_{G_3,T_1}) = avg(s_{G_4,T_1}) = \frac{1}{2}(s_{G_3,T_1} + s_{G_4,T_1})$. The result suggests that the gene "Yck2" should be annotated with the GO term "bud neck" and this is indeed a correct annotation according to the SGD database. For T_2 "cell budding", we calculate $avg(s_{G_1,T_2})$, $avg(s_{G_2,T_2})$, $avg(s_{G_3,T_2})$

and $avg(s_{G_4,T_2})$ first, and then assign it to the gene of the highest average score.

4.3.2 Density Model 2

In Density Model 1, each GO component has the same weight. The next model considers the weight of individual components of GO terms. We use *tf-idf* (term frequency and inverse document frequency) values to represent the importance of a GO-component, and propose Density Model 2. The second association algorithm is similar to Density Model 1, but each GO-component is multiplied by its *tf-idf* weight.

We formally explain the association algorithm of Density Model 2 as follows. We use the same symbols as Section 4.3.1. Besides, we define two more symbols.

$weight_{T_{j,k}}$: the *tf-idf* value of $T_{j,k}$ in GO and

$$weight_{T_{j,k}} = tf_{jk} \cdot \log_2 \frac{N}{n_{jk}} .$$ Where

tf_{jk} = frequency of the GO-component $T_{j,k}$ in T_j,
N = number of GO terms in the GO ontology, and
n_{jk} = number of GO terms where GO-component $T_{j,k}$ occurs at least once.

Then, the score between G_i and T_j with *tf-idf* weights is $\hat{s}_{G_i,T_j} = \frac{1}{c_j} \sum_{k=1}^{c_j} \frac{weight_{T_{j,k}}}{w_{G_i,T_{j,k}}}$. The formula to compute the average of all genes' occurrences identifying the same gene G_i, $avg(\hat{s}_{G_i,T_j})$, is the same as before except s_{G_i,T_j} is replaced with $\hat{s}_{G_i,T_j}$, so that

$$avg(\hat{s}_{G_i,T_j}) = \frac{1}{m} \sum_{l=1}^{m} \hat{s}_{G_{i_l},T_j} \quad \text{for all} \quad G_{i_l} = G_i \quad \text{and}$$

there are m occurrences with the same name G_i. Finally, a gene G_p with the highest value of $avg(\hat{s}_{G_p,T_j})$ will be associated with T_j.

We apply this algorithm to the example in Figure 4 again. Suppose the weights of "bud" and "neck" are w_1 and w_2. The score considering *tf-idf* values between G_1 and T_1 is $\hat{s}_{G_1,T_1} = \frac{1}{9}(\frac{w_1}{3} + \frac{w_2}{2} + \frac{w_1}{19} + \frac{w_1}{48} + \frac{w_2}{49} + \frac{w_1}{136} + \frac{w_2}{137} + \frac{w_2}{161} + \frac{w_1}{174})$. And $avg(\hat{s}_{G_1,T_1})$ is equal to $\hat{s}_{G_1,T_1}$. We then compute the values of $avg(\hat{s}_{G_2,T_1})$, $avg(\hat{s}_{G_3,T_1})$, and $avg(\hat{s}_{G_4,T_1})$ in a similar way. At last, we select the gene with the highest score. The association between T_2 "cell budding" and a gene is made in a similar way.

5 Experiment Results

5.1 Evaluation Metrics

We use the standard precision and recall evaluation measures. Precision and recall are defined as follows.

$$\text{Precision} = \frac{TP}{TP + FP}, \text{ and Recall} = \frac{TP}{TP + FN}$$

where *TP* is the number of true positives, *FP* is the number of false positives, and *FN* is the number of false negatives. In this experiment, true positives are the correct GO annotations proposed by our system. False positives are the incorrect GO annotations proposed by our system. False negatives are GO annotations in the answer key which our system did not propose.

However, the standard precision measurement is not representative of the performance of the system because the answer key is incomplete. The GO annotation files provided by the GO website do not contain every single correct GO annotation for all of the 4,479 articles in our corpus. This is because the GO website obtains its GO annotation files from different databases and these databases often specialize in different areas. For example, WormBase, one of the databases that the GO website gets its GO annotations from, focuses on Caenorhabditis elegans genes and gene products.[10] Therefore, annotations provided by WormBase may miss non-Caenorhabditis elegans genes. To account for this, we also provide an alternative precision measurement which assumes that a list of genes of interest for each article is given. This precision measurement ignores the proposed GO annotations which do not refer to one of the genes in the genes-of-interest list. For each article, we define its genes-of-interest list to be all the genes mentioned in the answer key's GO annotations which refer to that particular article. We call this precision measurement "Known Gene Precision" in Section 5.2. In a similar vein, we also define two other precision measurements, where one assumes that a list of GO terms of interest is given, and the other assumes that both a list of GO terms of interest and a list of genes of interest are given. The

---

[10] http://www.wormbase.org/

	TP	FP	FN	Recall	Precision	Known Gene Precision	Known GO Precision	Known GO-Gene Precision
Baseline	3,826	10,872,725	11,015	25.78%	0.04%	0.87%	3.37%	81.80%
System GO	1,631	1,521,440	13,210	10.99%	0.11%	1.37%	7.69%	87.08%
System Gene	469	442,974	14,372	03.16%	0.11%	1.41%	7.62%	84.77%
System GO/Gene	318	234,872	14,523	02.14%	0.14%	1.55%	8.98%	87.15%

Table 1. The Experimental Results of Density Model 1

	TP	FP	FN	Recall	Precision	Known Gene Precision	Known GO Precision	Known GO-Gene Precision
Baseline	3,826	10,872,725	11,015	25.78%	0.04%	0.87%	3.37%	81.80%
System GO	1,645	1,523,262	13,196	11.08%	0.11%	1.37%	7.75%	86.99%
System Gene	600	449,399	14,241	4.04%	0.13%	1.74%	7.72%	85.59%
System GO/Gene	385	237,605	14,456	2.59%	0.16%	1.83%	9.14%	87.73%

Table 2. The Experimental Results of Density Model 2

former will be named as "Known GO Precision" and the latter will be named as "Known GO-Gene Precision" in the following section.

5.2 Results and Discussion

The TP (true positive), FP (false positive), FN (false negative) values of Density Models 1 and 2 are shown in the left part of Tables 1 and 2, respectively. There is no TN (true negative) value, because the answer key does not contain false instances.

For the baseline, we proposed every single pair of GO term and gene names appearing in the same paragraph. The baseline provides an upper bound for the recall value. The "System GO" row shows the performance of our system as described in Section 4.3. For the "System Gene" experiment, we assigned one GO term to every gene, instead of the other way around. The same density-based method was used, except that the role of genes and GO-components were switched. That is, for each unique gene appearing in a paragraph, we associated exactly one GO term with it. For the "System GO/Gene" experiment, we proposed only the GO annotations that appear in both "System GO" and "System Gene" experiments. In other words, the set of GO annotations proposed in the "System GO/Gene" experiment is the intersection of the sets of annotations proposed in the "System GO" and "System Gene" experiments.

For the example mentioned in Figure 4, the baseline method would return 6 annotations because there are 3 distinct genes and 2 distinct GO terms. The "System GO" method will produce 2 annotations, each with a unique GO

term. The "System Gene" method will propose 3 annotations, because there are 3 different genes. Finally, the "System GO/Gene" method will generate at most 2 annotations, because the number of distinct GO terms limits the size of the intersection to the maximum of 2.

Experimental results under different precision metrics are shown in the right part of Tables 1 and 2. The "Precision" column shows the precision values as defined in the conventional way. The precision values in the "Known Gene Precision" column are obtained by assuming that the genes of interest are given. Similarly, values in the "Known GO Precision" column are computed assuming that the GO terms of interest are provided. For the "Known GO-Gene Precision" values, it is assumed that both the GO terms and genes of interest are given.

The experimental results show that recall rates decrease when the annotation conditions become stricter. For instance, the Density Model 1's baseline recall value of 25.78% drops to 2.14% when the "System GO/Gene" method is used. It is expected, because stricter conditions would filter out correct annotations where the GO term and gene are not close to each other. The recall rate of "System GO" is higher than "System Gene" in both models. This implies that the appearance of a GO term is a better indicator for the presence of a GO annotation than the appearance of a gene.

In Tables 1 and 2, the rank of different precision measurements, ordered from the lowest to the highest, is "Precision", "Known Gene Precision", "Known GO Precision" and "Known GO-Gene Precision". It tells us the relative difficulties of different annotation tasks.

For example, knowing GO terms of interest makes GO annotation easier than knowing genes of interest.

The high precision of "Known GO-Gene Precision" indicates that the word proximity relationship between Gene and GO terms really works. It shows that word proximity is a good feature for GO annotation.

Moreover, the performance of Density Model 2 is better than Density Model 1. It shows *tf-idf* values of GO-components also play an important role in GO annotation.

6   Conclusion

This paper uses the word proximity relationship between genes and GO terms to do GO annotation. We proposed an automatic way to assign a GO term to a gene based on the full-text documents. We made different experiments on GO annotation, including (1) proposing all pairs of gene-GO term, (2) assigning one gene to every GO term, (3) assigning one GO term to every gene, and (4) making the intersection of (2) and (3). We applied different precision metrics to evaluate the results. We also built two density models to explore the influence of GO-component's importance factor. The rank of experimental results using different precision metrics tells us the relative difficulties of different annotation tasks. The high performance of "Known GO-Gene Precision" reveals the word proximity relationship is a good feature for GO annotation. Furthermore, the *tf-idf* weight is also an important feature.

The preliminary experiments have promising results which will be helpful for the annotation task. There is still room for improvement. For improving the accuracy of GO term recognition, we can use the information such as word semantics or co-occurrence words. We also try to find other relationship between genes and GO terms and it's on-going. Combining word semantics to boost the performance is other feasible directions. Furthermore, combining other approaches with ours may increase the performance and that will be our future work.

Acknowledgements

References

Ashburner, M. and Lewis, S. (2002) On Ontologies for Biologists: the Gene Ontology—Untangling the Web, *Novartis Found Symposium*, **247**, 66–80.

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.

Ball, C.A., Dolinski, K. Dwight, S.S., Harris, M.A., Issel-Tarver, L., Kasarskis, A. *et al.* (2000) Integrating Functional Genomic Information into the Saccharomyces Genome Database, *Nucleic Acids Research*, **28**:1, 77-80.

BioTagger. http://www.seas.upenn.edu/~ryantm/ software /BioTagger/

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.

Flybase. http://flybase.bio.indiana.edu/

Hou, W.J., Lee C., Lin, K.H.Y. and Chen, H.H. (2005) A Relevance Detection Approach to Gene Annotation, *Proceedings of the first International Symposium on Semantic Mining in Biomedicine*, 15-24.

Joslyn, C.A., Mniszewski, S.M., Fulmer, A. and Heaton, G. (2004) The Gene Ontology Categorizer, *Bioinformatics*, **20(Suppl. 1)**, i169-i177.

Liu, H., Wu, C. and Friedman, C. (2004) BioTagger: A Biological Entity Tagging System, *BioCreative Workshop 2004 handout*, http://www.pdg.cnb. uam.es/BioLink/workshop_BioCreative_04/handout/pdf /Liu_Hongfang_BioTagger.pdf.

MGI. http://www.informatics.jax.org/

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.

Porter, M.F. (1980) An Algorithm for Suffix Stripping, *Program*, **14(3)**, 130−137.

Porter's stemmer. http://www.tartarus.org/~martin/ PorterStemmer/

Ray, S. and Craven, M. (2005) Learning Statistical Models for Annotating Proteins with Function Information Using Biomedical Text, *BMC Bioinformatics*, **6(Suppl. 1)**:S18.

SGD. http://www.yeastgenome.org/

The Gene Ontology Consortium (2001) Creating the Gene Ontology Resource: Design and Implementation, *Genome Research*, **11**, 1425–1433.

TREC Genomics Track. http://ir.ohsu.edu /genomics/

Verspoor, K., Cohn, J., Joslyn, C., Mniszewski, S., Rechtsteiner, A., Rocha, L.M. and Simas, T. (2005) Protein Annotation as Term Categorization in the Gene Ontology, *BMC Bioinformatics*, **6(Suppl. 1)**:S20.

WormBase. http://www.wormbase.org