Abstract

Gene Ontology (GO) is developed to provide standard vocabularies of gene products in different databases. The process of annotating GO terms to genes requires curators to read through lengthy articles. Methods for speeding up or automating the annotation process are thus of great importance. We propose a GO annotation approach using full-text biomedical documents for directing more relevant papers to curators. This system explores word density and gravitation relationships between genes and GO terms. Different density and gravitation models are built and several evaluation criteria are employed to assess the effects of the proposed methods.

Keywords: Gene Ontology, GO annotation, gravitation model, density model

1 Introduction

Over the past years, a vast and growing amount of gene data stored in various databases has made it necessary to integrate gene information spread over several different databases. Gene Ontology (GO) [2, 15] provides standard vocabularies for gene products in different databases. GO includes three sub-ontology classifications that describe three semantic types of concepts. These sub-ontologies represent different genomic characteristic categories, which are described by GO terms. Each GO term is associated with a unique ID called a “GO ID.” Several databases (e.g. SGD [23], Flybase [16], and MGI [18]) specializing in different organisms annotate their gene products with GO terms, and provide references that indicate evidence available to support the annotations. The annotation process requires curators to read through articles tediously. Methods for hastening or automating the annotation process to meet the large volume of literature are highly sought after.

Automatic GO annotations contain two types of labeling: document labeling [10, 13] and gene labeling [5, 6, 7, 12]. Document labeling assigns GO terms to documents without denoting relevant genes, while gene labeling explicitly associates genes with GO terms.


Except Ruch, all researchers mentioned above used small data collections, for example, 50 genes in [5] and 21 GO terms in [12]. Although Ruch considered most GO terms, his work belonged to document labeling. Most previous studies used information from the article title, abstract and MeSH
terms only. Therefore, introducing full-text articles and reducing noise resulting from full texts is another challenge in this study. In contrast to previous related works, the current study conducts large-scale gene annotation at the level of gene-labeling using large collections of GO terms and full-text documents.

In this paper, we propose two GO annotation features: the density and gravitation functions. The “density” feature is used to measure the proximity of genes and GO terms in a paragraph. The “gravitation” feature is adapted to simulate the intensity of genes and GO terms on physical domain. We want to investigate whether the physical phenomena such as Newton’s gravitation law and density distribution can help with the GO annotation. By observation, if only one gene and GO term appears in the same paragraph, they are considered to be associated with each other. If more than one gene is found, the gene closer to the GO term or more gravitation between a gene and GO terms is preferred. Consider the GO term “endocytosis” in the literature with PMID 9045618 as follows:

... Re-expression of <GENE>tuberin</GENE> in <GENE>TSC2</GENE> mutant cells reduced the rate of fluid-phase <GO>endocytosis</GO>. These findings suggest that <GENE>tuberin</GENE> functions as a Rab5GAP in vivo to negatively regulate Rab5-GTP activity in <GO>endocytosis</GO>.

There are two genes and one GO term with three gene-GO_term occurrences in the above paragraph. The gene nearest the GO term “endocytosis” is “tuberin,” which is annotated with “endocytosis” in UniProt [1]. The “TSC2” term is farther from the GO term, and “TSC2” is not annotated with “endocytosis.” This indicates that associating GO terms with the nearest gene is a reasonable postulation. Another interesting issue is how to find a selected number of features such that the binding of “tuberin” and “endocytosis” is tighter than “TSC2” and “endocytosis.”

A 3-tuple representation describes the relationship between PMID, GO terms and genes in this study. For example, this is a triple of <PMID, GO ID, GENE> where “PMID” represents the PubMed identifier, “GO ID” represents the GO category ID, and “GENE” represents the gene name.

The rest of this paper is organized as follows. Section 2 provides an overview of the system architecture. Section 3 depicts the experimental corpus construction. Section 4 presents details of the proposed methods, and Section 5 discusses experimental results. This section also introduces evaluation metrics. Finally, we make a number of conclusions.

2 System Overview

Figure 1 shows the overall architecture of the proposed system. First, we preprocessed each full-text article in the corpus. We made use of some specific resources in the biomedical domain (e.g. BioTagger [9] and Gene Ontology) and some natural language processing resources (e.g. a stop-word list and Porter’s stemmer [11]) to preprocess the corpus. Next, we got a set of articles with tagged gene names and tagged GO terms. Then an algorithm based on word density or gravitation relationship annotated genes with GO terms. Finally, a 3-tuple of <PMID, GO ID, GENE>, which specified a gene GENE, was annotated with some GO terms in a biomedical article PMID.

3 Corpus Consideration

In this stage, we downloaded all the GO annotation files from the GO website [17]. The website shows the information of gene products annotated to GO. It includes the databases contributing to this association, the unique ID in databases, the GO ID, the literature references (such as PMID), and other related information. We searched for each entry’s corresponding biomedical article using Entrez PubMed [19]. Then we downloaded the free online full-text articles if they existed. We were able to retrieve 10,054 full-text articles. As we mentioned before, several databases specializing in
different organisms annotate their gene products with GO terms. Therefore, the corpus was composed of articles covering different organisms.

Figure 1: System architecture.

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. The focus of this study was not to recognize all the different names of a gene, so we filtered out the articles which did not contain the gene names specified in the GO annotation file. Each article was examined and we kept it if at least one entry in the article's GO annotation file also referred to a gene that appeared in the article. After filtering, 4,479 articles remained. We also removed entries in the GO annotation files which did not refer to one of these 4,479 articles or did not refer to a gene name that appeared in one of these 4,479 articles. In summary, our final corpus comprised 4,479 full-text articles, which contained a total of 15,566 annotations.

4 Methods

Our annotation procedure for each article consisted of gene name tagging, GO term tagging, and GO-term and gene-name association. The untagged example for the following annotation procedure is illustrated in Figure 2.

4.1 Gene Name Tagging

We used BioTagger to identify every time a gene is mentioned in an article [20]. BioTagger is a biological entity tagging system capable of recognizing gene names, genomic variations in cancers, and
It is notable that no EMB-9 or LET-2 reporter activity was seen in pharyngeal, intestinal, or hypodermal cells of embryos, larvae, or adults. This result is consistent with the fact that collagenIV anti-type antisera did not show intracellular stain in these tissues. However, the pharynx and intestine are covered by basement membranes that are stained by the antisera collagen IV anti-type. Although the pharyngeal and intestinal basement membranes contain EMB-9 and LET-2, the genes encoding them are not expressed in these tissues.

Figure 2: Untagged paragraph with PMID 9166416.

malignancy types in cancers. BioTagger has a 77% precision rate and a 96% recall rate for the yeast.

4.2 GO Term Tagging

We used word matching to identify GO terms. First, we used PubMed's stop-word list to remove all the stop-words from every GO term. We then stemmed the GO terms with Porter's stemmer [21]. When tagging GO terms, we treated each paragraph of an article as an independent unit because different paragraphs usually focus on different GO terms. 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. Figure 3 shows the tagging results of the paragraph in Figure 2. There are two different genes and three distinct GO terms in Figure 3. The matching words with the GO term “basement membrane” (GO:0005604) are in boldface.

GO:0001775: cell activation
GO:0005604: basement membrane
GO:0005622: intracellular

Figure 3: Example of gene name tagging and GO term tagging. The upper part is the result of gene and GO term tagging. The lower part lists the GO terms indicating as T1, T2 and T3, respectively.

4.3 GO to Gene Association

Figure 4 presents the basic concept of our GO-to-Gene association algorithm. In Figure 4, it illustrates one paragraph where there are two genes and four GO-components. Two genes, g1 and g2, and four GO-components may be the same or different in a paragraph. Without loss of generality, we assume there is only one GO term in this paragraph, and there is a link between each gene and each GO-component. We then calculate the scores between them. The gene closest to the GO term should be associated with the GO term. In other words, the shorter the distance between a gene and a GO-component is, the higher the score. More important GO-components have higher scores. This model is similar to a density model where the gene with the highest density (i.e. most tightly surrounded by
GO-components) will be selected. We designed a density model to explore the relationship between word distance and the GO-component’s importance. This model will be described in Section 4.3.1.

Furthermore, we examine GO annotation from the perpectives of physics. Inspired by Newton’s theory of gravitation, Shi et al. [14] proposed a gravitation-based model for information retrieval (IR). They mapped IR concepts to physics concepts (mass, distance, radius, etc.). We designed experiments such as gravitation models to study the effects based on the gravitation, which will be explained in Section 4.3.2.

4.3.1 Density Model

We associated one gene with every GO term appearing in a paragraph. We also considered the weight of individual GO term components except for the effect of the distance between GO-component and genes. We used tf-idf (term frequency and inverse document frequency) values in IR to represent the importance of a GO-component. The GO-to-Gene association algorithm is stated informally as follows.

For each occurrence of a gene in a paragraph, we compute the distance between the gene occurrence and GO-component, and determine the weight of individual GO term components. The shorter the distance is, the higher the score, and the higher the tf-idf is, the higher the score, too. We then average the scores of the gene and each GO-component. Finally, we average all gene occurrence scores. The gene with the highest score is associated with the GO term. To describe the algorithm more formally, we define the symbols as follows:

\[ G_i : \text{the } i\text{-th gene occurrence in a given paragraph}, \]
\[ T_j : \text{the GO term with the } j\text{-th unique GO ID}, \]
\[ T_{j,k} : \text{the } k\text{-th occurrence of a GO term } T_j \text{'s GO-component in a given paragraph}, \]
\[ w_{G_i,T_{j,k}} : \text{total number of words between gene } G_i \text{ and GO-component } T_{j,k}, \]
\[ c_j : \text{total occurrences of } T_j \text{'s GO-components in a given paragraph, and} \]
\[ tfidf_{T_{j,k}} : \text{the } tf-idf \text{ value of } T_{j,k} \text{ in GO and} \]
\[ tfidf_{T_{j,k}} = tf_{jk} \cdot \log_2(\frac{N}{n_{jk}}). \]
\[ tf_{jk} = \text{frequency of the GO-component } T_{j,k} \text{ in } T_j, \]
\[ N = \text{number of GO terms in the GO ontology, and} \]
\[ n_{jk} = \text{number of GO terms where GO-component } T_{j,k} \text{ occurs at least once}. \]

For example, Table 1 shows the values of \( G_i, T_j, T_{j,k}, w_{G_i,T_{j,k}} \) and \( c_j \) of Figure 3. For simplicity, we only list four instances of \( w_{G_i,T_{j,k}} \) out of 28 instances, which are formed by four gene occurrences and seven GO-component occurrences.

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

For a certain GO term \( T_j \), the average score of all gene occurrences identifying the same gene \( G_i \) is \( \text{avg}(s_{G_i,T_j}) = \frac{1}{m} \sum_{i=1}^{m} s_{G_i,T_j} \), for all \( G_i = G_i \) and there are \( m \) occurrences with the same name \( G_i \).
Finally, a gene $G_p$ with the highest score of $\text{avg}(s_{G_p,T_j})$ will be associated with $T_j$.

We apply our algorithm to the example in Figure 3. For simplicity, we only demonstrate how $T_2$, “basement membrane,” associates with a gene. The associations between a gene and $T_1$ and $T_3$ are not shown. First of all, genes “emb-9” and “let-2” appear twice in the paragraph. $G_1$, “emb-9,” has distances of 43, 44, 56, and 57, to the four occurrences of $T_2$’s GO-components. Suppose the weights of “basement” and “membrane” are $w_1$ and $w_2$. The association score between $G_1$ and $T_2$ is: $s_{G_1,T_2}$. Then $\text{avg}(s_{G_1,T_2})$ is equal to $\frac{1}{2}(s_{G_1,T_2} + s_{G_3,T_2})$, which is also equal to $\text{avg}(s_{G_3,T_2})$. Moreover, we compute $s_{G_2,T_2}$ and $s_{G_4,T_2}$ where $G_2$ = “let-2”. We also calculate $\text{avg}(s_{G_2,T_2}) = \text{avg}(s_{G_4,T_2}) = \frac{1}{2}(s_{G_2,T_2} + s_{G_4,T_2})$. Experimental results suggest that the “emb-9” gene should be annotated with the GO term “basement membrane” and this is indeed a correct annotation according to the WormBase database [22].

4.3.2 Gravitation Model

The gravitation model attempts to deal with the GO annotation problem by using a physical framework. We map GO annotation concepts to physical concepts as follows. Since a tagged word in the text is either a gene or a GO-component, we consider a basic particle in Newton’s gravitation law to be a gene or a GO-component in GO annotation. For two particles $G_i$ and $T_{j,k}$ with mass $m_{G_i}$ and $m_{T_{j,k}}$, respectively, the gravitational force between them is $F_{G_i,T_{j,k}} = \frac{G m_{G_i} m_{T_{j,k}}}{d^2}$, where $d$ is the distance between the two particles and $G$ is a constant.

We establish different gravitation models (GMs) according to the variations of $m_{G_i}$ and $m_{T_{j,k}}$. In the model, $m_{G_i}$ is affected by term frequencies in paragraphs and documents. If a gene appears frequently in a paragraph and in the whole document, it should have a higher mass because the gene is more important to the document. Thus, we let $m_{G_i} = \frac{p_{t,f} G_i}{\text{idf}_{G_i}}$, where $p_{t,f} G_i$ is the frequency of $G_i$ in the paragraph, and $\text{idf}_{G_i}$ is the idf value of $G_i$. Here, we use the reciprocal of idf because frequent appearance in a document is preferred. We also use the $t,f-idf$ weight, i.e. $t_{f-idf} T_{j,k}$ for $T_{j,k}$, in Section 4.3.1 as the value of $m_{T_{j,k}}$.

The gravitational force between $G_i$ and $T_{j,k}$ is $F_{G_i,T_{j,k}} = m_{G_i} m_{T_{j,k}}/d^2 = (p_{t,f} G_i/\text{idf}_{G_i}) \times t_{f-idf} T_{j,k}/d^2$. The gravitational force between $G_i$ and $T_j$ is the sum of the forces between $G_i$ and each $T_{j,k}$: $F_{G_i,T_j} = \sum_k F_{G_i,T_{j,k}}$. The formula to compute the average of all gene occurrences identifying the same gene $G_i$, $\text{avg}(F_{G_i,T_j})$, is the same as before, so that $\text{avg}(F_{G_i,T_j}) = \frac{1}{m} \sum_{i=1}^m F_{G_i,T_{j,k}}$ for all $G_i = G_i$ and there are $m$ occurrences with the same name $G_i$. Finally, a gene $G_p$ with the highest value of $\text{avg}(F_{G_p,T_j})$ will be associated with $T_j$.

We call the above gravitation model GM1. Then we modify the mass of gene in GM2, GM3 and GM4. In GM2, we highlight the importance of the idf value of the gene. In GM3, we increase the gene’s mass. Whereas in GM4, we decrease the gene’s mass. Table 2 lists formulas for $F_{G_i,T_{j,k}}$.

### Table 1: Instances of each symbol in Figure 3.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Instance</th>
<th>Symbol</th>
<th>Instance</th>
<th>Symbol</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$</td>
<td>emb-9</td>
<td>$T_{1,1}$</td>
<td>Activity</td>
<td>$w_{G_1,T_{1,1}}$</td>
<td>3</td>
</tr>
<tr>
<td>$G_2$</td>
<td>let-2</td>
<td>$T_{1,2}$</td>
<td>Cells</td>
<td>$w_{G_1,T_{1,2}}$</td>
<td>11</td>
</tr>
<tr>
<td>$G_3$</td>
<td>EMB-9</td>
<td>$T_{2,1}$</td>
<td>Basement</td>
<td>$w_{G_2,T_{2,1}}$</td>
<td>1</td>
</tr>
<tr>
<td>$G_4$</td>
<td>LET-2</td>
<td>$T_{2,2}$</td>
<td>Membranes</td>
<td>$w_{G_2,T_{2,2}}$</td>
<td>9</td>
</tr>
<tr>
<td>$T_1$</td>
<td>cell activation</td>
<td>$T_{2,3}$</td>
<td>Basement</td>
<td>$c_1$</td>
<td>2</td>
</tr>
<tr>
<td>$T_2$</td>
<td>basement membrane</td>
<td>$T_{2,4}$</td>
<td>Membranes</td>
<td>$c_2$</td>
<td>4</td>
</tr>
<tr>
<td>$T_3$</td>
<td>intracellular</td>
<td>$T_{3,1}$</td>
<td>Intracellular</td>
<td>$c_3$</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2: Formulas of gravitational force in different gravitation models.

<table>
<thead>
<tr>
<th>Gravitation Model</th>
<th>Formula for $F_{G_i,T_{j,k}}$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM1</td>
<td>$(pt f_{G_i}/idf_{G_i}) \times tfidf_{T_{j,k}}/d^2$</td>
<td>The original gravitational force.</td>
</tr>
<tr>
<td>GM2</td>
<td>$(pt f_{G_i}/(idf_{G_i})^2) \times tfidf_{T_{j,k}}/d^2$</td>
<td>Emphasis of the importance of the $idf$ value of the gene.</td>
</tr>
<tr>
<td>GM3</td>
<td>$(pt f_{G_i}/(idf_{G_i})^2) \times tfidf_{T_{j,k}}/d^2$</td>
<td>Emphasis of the mass of the gene.</td>
</tr>
<tr>
<td>GM4</td>
<td>$(\sqrt{pt f_{G_i}/idf_{G_i}} \times tfidf_{T_{j,k}}/d^2$</td>
<td>Emphasis of the mass of the GO-component.</td>
</tr>
</tbody>
</table>

5 Experiment Results

5.1 Evaluation Metrics

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

\[
\text{Precision} = TP/(TP + FP), \quad \text{and} \\
\text{Recall} = TP/(TP + FN)
\]

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. Alternately, false positives are the incorrect annotations proposed by our system. False negatives are GO annotations in the answer key, which is composed of GO-gene association derived from the GO website, but our system does 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 all the correct GO annotations 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. To account for this, we also used an alternative precision measurement, which assumed that a list of genes of interest for each article was shown in our previous research [8]. It 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 a list containing 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. Similarly, we define two other precision measurements in the following section. 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 former is called “Known GO Precision” and the latter is called “Known GO-Gene Precision.” The set up of “Known Gene Precision” meets the needs of researchers who are often interested in particular genes. The measure of “Known GO Precision” is useful for researchers who want to look for genes with similar relationships to given GO terms. In addition, “Known GO-Gene Precision” can evaluate the GO-Gene relation independent of the errors that occur when tagging GO and gene names.

5.2 Results and Discussion

Tables 3 and 4 show the experimental results of density model and gravitation model, respectively. For the baseline, we proposed every single pair of GO term and gene names that appeared 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. The “System Gene” experiment assigned one GO term to every gene, instead of one gene to every GO term. The same density- and gravitation-based methods were used, except that the role of genes and GO-components were switched. That is,
we associated exactly one GO term with each unique gene appearing in a paragraph. For the “System GO/Gene” experiment, we proposed only the GO annotations that appeared in both “System GO” and “System Gene” experiments.

Table 3: Experimental results of density models.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall</th>
<th>Precision</th>
<th>Known Gene Precision</th>
<th>Known GO Precision</th>
<th>Known GO-Gene Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.78%</td>
<td>0.04%</td>
<td>0.87%</td>
<td>3.37%</td>
<td>81.80%</td>
</tr>
<tr>
<td>System GO</td>
<td>11.08%</td>
<td>0.11%</td>
<td>1.37%</td>
<td>7.75%</td>
<td>86.99%</td>
</tr>
<tr>
<td>System Gene</td>
<td>4.04%</td>
<td>0.13%</td>
<td>1.74%</td>
<td>7.72%</td>
<td>85.59%</td>
</tr>
<tr>
<td>System GO/Gene</td>
<td>2.59%</td>
<td>0.16%</td>
<td>1.83%</td>
<td>9.14%</td>
<td>87.73%</td>
</tr>
</tbody>
</table>

Table 4: Experimental results of gravitation models.

<table>
<thead>
<tr>
<th>Models</th>
<th>Recall</th>
<th>Precision</th>
<th>Known Gene Precision</th>
<th>Known GO Precision</th>
<th>Known GO-Gene Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System GO</td>
<td>13.82%</td>
<td>0.15%</td>
<td>1.18%</td>
<td>11.61%</td>
<td>87.68%</td>
</tr>
<tr>
<td>System Gene</td>
<td>3.74%</td>
<td>0.12%</td>
<td>1.65%</td>
<td>7.77%</td>
<td>85.80%</td>
</tr>
<tr>
<td>System GO/Gene</td>
<td>1.27%</td>
<td>0.26%</td>
<td>1.35%</td>
<td>18.27%</td>
<td>87.32%</td>
</tr>
<tr>
<td>GM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System GO</td>
<td>14.43%</td>
<td>0.16%</td>
<td>1.11%</td>
<td>13.08%</td>
<td>87.89%</td>
</tr>
<tr>
<td>System Gene</td>
<td>3.74%</td>
<td>0.12%</td>
<td>1.65%</td>
<td>7.77%</td>
<td>85.80%</td>
</tr>
<tr>
<td>System GO/Gene</td>
<td>1.43%</td>
<td>0.28%</td>
<td>1.33%</td>
<td>19.88%</td>
<td>88.60%</td>
</tr>
<tr>
<td>GM3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System GO</td>
<td>14.73%</td>
<td>0.16%</td>
<td>1.10%</td>
<td>13.56%</td>
<td>88.36%</td>
</tr>
<tr>
<td>System Gene</td>
<td>3.74%</td>
<td>0.12%</td>
<td>1.65%</td>
<td>7.77%</td>
<td>85.80%</td>
</tr>
<tr>
<td>System GO/Gene</td>
<td>1.45%</td>
<td>0.27%</td>
<td>1.29%</td>
<td>19.38%</td>
<td>88.42%</td>
</tr>
<tr>
<td>GM4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System GO</td>
<td>12.85%</td>
<td>0.13%</td>
<td>1.24%</td>
<td>10.02%</td>
<td>87.19%</td>
</tr>
<tr>
<td>System Gene</td>
<td>3.74%</td>
<td>0.12%</td>
<td>1.65%</td>
<td>7.77%</td>
<td>85.80%</td>
</tr>
<tr>
<td>System GO/Gene</td>
<td>1.05%</td>
<td>0.22%</td>
<td>1.28%</td>
<td>16.02%</td>
<td>87.50%</td>
</tr>
</tbody>
</table>

In Tables 3 and 4, the “Precision” column shows conventionally defined precision. 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, we assume that both the GO terms and genes of interest are given.

Results show that recall rates decrease when annotation conditions become stricter. For instance, when using the “System GO/Gene” in the density model, the baseline recall value of 25.78% drops to 2.59%. This is expected, because stricter conditions would filter out correct annotations where GO term and gene are not close to each other. The “System GO” recall rate is higher than the “System Gene” rate in both models. This implies that the appearance of a GO term is a better indicator of the presence of a GO annotation than the appearance of a gene.

The rank of different precision measurements, from the lowest to the highest, is “Precision,” “Known Gene Precision,” “Known GO Precision,” and “Known GO-Gene Precision.” The ranking...
of different precisions indicates the relative difficulties of different annotation tasks. For example, knowing GO terms of interest makes GO annotation easier than knowing genes of interest.

The "Known GO-Gene Precision" rates in Table 3 indicate that the word density between genes and GO terms is significant. The gravitation model performance is slightly better than the density model performance in most cases. This indicates that word density and gravitational force between genes and GO terms both play an important role in GO annotation.

We analyzed the "Known GO-Gene Precision" shown in Table 4. The better performance of GM2 over GM1 shows that the frequency of genes in the document is more significant than the frequency in the paragraph. A gene which appears frequently in a certain paragraph but not in the other paragraphs is often used to compare the target that the paper focused on. However, because the "Known GO-Gene Precision" of GM3 is better than that of "GM4," genes should be given more attention in the gravitation models. In this case, a gene is often surrounded by several GO-components, so that the gene should enhance its mass to attract the scattered GO-components.

Figure 5 lists the number of gene and GO tags that were identified for each of the gravitation models. It shows the scale of our experiments and it is stable for different models.

![Figure 5: Number of gene/GO tags in different gravitation models.](image)

<table>
<thead>
<tr>
<th>Method</th>
<th>Evaluation metrics and performance</th>
<th>Labeling method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perez et al. [10]</td>
<td>8% recall rate and a 67% precision rate</td>
<td>Document labeling</td>
</tr>
<tr>
<td>Ruch [13]</td>
<td>average precision rate &lt;10%</td>
<td>Document labeling</td>
</tr>
<tr>
<td>Homayouni et al. [5]</td>
<td>average precision rate 10%-72%</td>
<td>Gene labeling</td>
</tr>
<tr>
<td>Hou et al. [6]</td>
<td>78% recall rates and 66% precision rates at distance 12</td>
<td>Gene labeling</td>
</tr>
<tr>
<td>Kim and Park [7]</td>
<td>89.7% precision rate</td>
<td>Gene labeling</td>
</tr>
<tr>
<td>Raychaudhuri et al. [12]</td>
<td>72% accuracy</td>
<td>Gene labeling</td>
</tr>
<tr>
<td>GM4 in System GO/Gene</td>
<td>88.60% Known GO-Gene Precision</td>
<td>Document labeling</td>
</tr>
</tbody>
</table>

Section 1 presents some related studies. They are briefly listed with our partial result in Table 5. Varying test sets and evaluation criteria did not provide enough evidence to conclude which methods performed better. In the scalability viewpoint of GO terms, we might compare Ruch's performance [13] to ours. Results are similar at first glance, but the experiments are different in two aspects: (1) we perform finer grain labeling, i.e. GO annotation at the level of gene labeling rather than document
GO Annotation by Density and Gravitation Models

labeling, and (2) our test corpus of 4,479 full-text documents containing a total of 15,566 annotations is larger than the BioCreative benchmark of 99 test articles with 1,227 annotations [3].

GO annotation is challenging and shows low performance for several reasons. First, GO has a large number of categories. Secondly, the nature of GO concepts is to express biological descriptions, not textual contents. Thirdly, using the precision metric for the hierarchical structure may incur information loss, so Perez et al. [10] proposed a new metric. Though our results did not show significant improvements, we proposed a novel approach, i.e. using density and gravitation models, and applied them to large-scale collections of GO terms and full-text testing documents, which is a real-life scenario that curators have to deal with when annotating GO.

6 Conclusion

This paper uses word density and gravitation relationships between genes and GO terms to perform GO annotation. We propose an automatic way to assign a GO term to a gene based on full-text articles. We performed different GO annotation experiments, 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) combining (2) and (3). We applied different precision metrics to evaluate the results and built a density model to explore the influence of the GO-component’s importance factor. We also built four gravitation models to explore the influence of the gravitational force between genes and GO terms. Experimental results using different precision metrics indicate the relative difficulties of different annotation tasks. The high performance of "Known GO-Gene Precision" reveals that word density and gravitation relationships are good GO annotation features.

Acknowledgments

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts, 94-2213-E-002-033 and 95-2752-E-001-001-PAE.

References


