Inf Retrieval (2008) 11:389–404
DOI 10.1007/s10791-008-9053-0

Combining gene sequence similarity and textual information for gene function annotation in the literature

Luo Si · Danni Yu · Daisuke Kihara · Yi Fang

Received: 4 September 2007 / Accepted: 25 January 2008 / Published online: 9 February 2008

Abstract Annotation of the functions of genes and proteins is an essential step in genome analysis. Information extraction techniques have been proposed to obtain the function information of genes and proteins in the biomedical literature. However, the performance of most information extraction techniques of function annotation in the biomedical literature is not satisfactory due to the large variability in the expression of concepts in the biomedical literature. This paper proposes a framework to improve the gene function annotation in the literature by considering both the textual information in the literature and the functions of genes with sequences similar to a target gene. The new framework collects multiple types of evidence as: (i) textual information about gene functions by matching keywords of the gene functions; (ii) gene function information from the known functions of genes with sequences similar to a target gene; and (iii) the prior probabilities of gene functions to be associated with an arbitrary gene by mining the known gene functions from curated databases. A supervised learning method is utilized to obtain the weights for combining the three types of evidence to assign appropriate Gene Ontology terms for target genes. Empirical studies on two testbeds demonstrate that the combination of sequence similarity scores, function prior probabilities and textual information improves the accuracy of gene function annotation in the literature. The *F*-measure scores obtained with the proposed framework are substantially higher than the scores of the solutions in prior research.

L. Si (✉)
Department of Computer Science and Statistics, Purdue University, West Lafayette, IN 47906, USA
e-mail: lsi@cs.purdue.edu

D. Yu
Department of Statistics, Purdue University, West Lafayette, IN 47906, USA
e-mail: dyu@purdue.edu

D. Kihara
Department of Biology and Computer Science, Purdue University, West Lafayette, IN 47906, USA
e-mail: dkihara@purdue.edu

Y. Fang
Department of Computer Science, Purdue University, West Lafayette, IN 47906, USA
e-mail: fangy@cs.purdue.edu

Keywords Biomedical literature mining · Gene function annotation · Combination of multiple evidence

1 Introduction

The rapid advancements of high-throughput methods for acquiring gene sequences provide a wealth of valuable biomedical knowledge about gene functions. Apart from curated databases that contain some gene function information (e.g., the EBI,[1] the MGI Mouse[2] or the UniProt[3] databases), a large body of biomedical knowledge of gene functions exists in the biomedical literature.

The number of published biomedical research papers, and the resulting underlying biomedical knowledge, are growing at an increasing rate. The PubMed/MEDLINE literature database at NIH, which contains more than 17 millions biomedical records,[4] can serve as an important information source for gene function annotation. The automatic process of function annotation from the literature is a very important task in two aspects: First, a large amount of "hidden" function information of genes, which is not explicitly available from curated databases, can be extracted from text literature. Second, the accurate and comprehensive links of gene functions to the biomedical literature aid biomedical researchers to validate identified gene functions through the context information (e.g., passages or abstracts), which is more informative than only the function identities from curated databases.

The task of gene function annotation from the literature is currently conducted manually by human curators (Camon et al. 2004), which is often an expensive and time-consuming process. Automated discovery of the gene functions has become very important because of the enormous amount of the biomedical literature published every year. Recently, many information extraction techniques have been proposed and developed (Blaschke 2005) for gene function annotation in the biomedical literature by assigning functions (more than 23,000 GO terms) in the Gene Ontology[5] (i.e., GO). Although substantial progress has been made to improve the accuracy of function annotation in the literature, empirical studies in the BioCreAtIvE challenge (Blaschke 2005) and other related research (e.g., Stoica and Hearst 2006) demonstrate that current systems are not yet able to produce satisfactory results. The performance of the results in the annotation subtask in the TREC Genomics task (Hersh et al. 2004) is higher since this is a simplified task, which only requires to annotate the three top GO hierarchies (i.e., biological process, molecular function and cellular component) in the Gene Ontology.

One approach of gene annotation in the literature is keyword-based with the focus on textual information (Hersh et al. 2004; Blaschke 2005; Cohen and Hersh 2005). However, keyword-based approaches often have limited power to address the large variability in the expression of concepts in the biomedical literature. When a GO term is actually associated with a gene in a piece of text, maybe only some of the tokens of the GO term appear in this piece of text. On the other hand, even if many tokens of a

[1] http://www.ebi.ac.uk/Databases/

[2] http://www.informatics.jax.org/

[3] http://www.pir.uniprot.org/

[4] http://www.nlm.nih.gov/pubs/techbull/ma07/ma07_technote.html#8

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

GO term appear in a piece of text with a gene, it may not be appropriate to annotate the gene with the GO term either because the GO term is used to describe another biomedical concept or because the GO term is too general for the gene. It is often difficult to design sophisticated keyword-based information extraction techniques to capture the large variability in the text data. Because of these limitations, the performance of most keyword-based information extraction techniques for gene function annotation in the literature is not satisfactory.

Some previous research has been proposed (Koike et al. 2004; Settles and Craven 2004; Couto et al. 2005; Ehrler and Ruch 2005; Krallinger et al. 2005; Ruch et al. 2005) to address the limitation of the keyword-based techniques. The previous research only focused on textual information. This paper proposes a framework to improve the accuracy of gene function annotation in the literature by considering multiple types of evidence as: textual information in the literature, functions of genes with high sequence similarity scores to the target gene and the prior probabilities of gene functions.

Textual information is collected by matching the tokens of a GO term and calculating their weights. For a GO term, our method attempts to identify the occurrences of each token in the GO term and aggregates the weights of individual tokens, which measure the importance of the tokens associated with the GO term. Similar approaches have been used in some previous research (Koike et al. 2004; Krallinger et al. 2005).

For a target gene, some gene function information is obtained from the known functions of genes with sequences similar to the target gene. Particularly, conventional BLAST searches (Altschul et al. 1990) are used to identify genes with sequences similar to the target gene. Direct evidence of gene function is obtained from the known functions of these similar genes. Furthermore, indirect evidence of gene function is obtained by considering the gene functions that are highly correlated with the known functions of the similar genes.

Moreover, when we consider annotating a target gene with a specific GO term in the literature, the prior probability of the GO term is also investigated. The prior probabilities of available GO terms are calculated based on the frequency of the GO terms occurring in the UniProt-SwissProt database (Wu et al. 2006).

This paper proposes a supervised learning method for combining the three types of evidence to assign appropriate GO terms to target genes. A set of useful features is extracted from the three types of evidence mentioned above. Specifically, a logistic regression model is estimated to obtain the weights of these features from a set of relatively small amount of training data. Finally, the logistic regression model can be utilized to find gene function annotations in the biomedical literature.

An extensive set of experiments has been conducted on both the EBI Human and the MGI Mouse datasets to demonstrate the advantages of the proposed framework for gene function annotation in the literature. Experimental results have shown that it is beneficial to incorporate other types of evidence beyond the textual information for gene function annotation in the literature. Furthermore, the proposed framework with all the three types of evidence has been shown to substantially improve the annotation accuracy (i.e., in F-measure, Rijsbergen 1979) than the accuracy of two solutions in prior research.

The next section discusses related work. Section 3 describes the new framework that considers multiple types of evidence with the supervised learning method. Section 4 explains our experimental methodology. Section 5 presents the experimental results and the corresponding discussion. Section 6 concludes.

2 Related work

Automatic gene function prediction is an important and popular research topic in the bioinformatics community (Hawkins and Kihara 2007). Different types of approaches have been proposed to utilize sequence-based gene information (e.g., Jensen et al. 2003), gene expression profiles (e.g., Eisen et al. 1998) and comparative genomics methods (e.g., Marcotte et al. 2000) for automatic gene function prediction.

Gene function annotation in the biomedical literature has a different goal than automatic gene function prediction. Gene function annotation in the biomedical literature identifies biomedical documents that verify the associations between gene functions with target genes. Automatic gene function prediction predicts new functions for genes/proteins.

Gene function annotation in the biomedical literature is a relatively new research topic (Hersh et al. 2004; Blaschke 2005). Most prior work utilized information extraction techniques that focus on the textual information of genes, gene functions and biomedical documents.

The Meke (Medical Knowledge Explorer) system (Chiang and Yu 2003, 2004) identifies common phrase patterns in the biomedical literature for describing gene functions. For example, one pattern could be "gene is associated with a function". Pattern mining techniques are utilized to find those common patterns. Finally, a Naïve Bayesian classifier is applied to predict the likelihood that a sentence describes a gene-function relation.

The Figo system (Couto et al. 2005) annotates a gene with a GO term in a document by considering the information content of a GO term, which is calculated as a function of the matched tokens in the document. Similarly, different GO terms are treated as different classes in Ehrler and Ruch (2005) and each biomedical document is assigned with different classes (i.e., GO terms) based on weighted word matching.

The text representation of a GO term is expanded in Ray and Craven (2005) by adding words that often co-occur with the tokens of the GO term. Furthermore, a statistical classifier is utilized to find gene function annotation by considering the original text representation of a GO term and the expanded text representation. The work in Verspoor et al. (2004) also expands the text representations of GO terms with words that have strong associations with the tokens of the GO terms. Furthermore, it uses a classifier (Joslyn et al. 2004) that considers the structure of Gene Ontology to find gene function annotation in the literature. The support vector machine technique has been used in Rice et al. (2005) for gene function annotation in the biomedical literature. The work in Raychaudhuri et al. (2002) compares several types of classifiers for assigning only 21 GO terms. The text classification approach for identifying GO terms is related with a large body of previous research for identifying Medical Subject Headline (MeSH) topics. Different techniques such as K nearest neighbor (KNN), Naïve Bayes and Support Vector Machine have been applied to identify MeSH topics in PubMed abstracts (Yang 1999; Joachims 1998; Sebastiani 1999; Kim and Wilbur 2005).

All the above work only considers textual information. Very limited research has been conducted to combine other types of evidence for gene function annotation in the biomedical literature. Stoica and Hearst (2006) proposes a nice approach to improve the accuracy of gene function annotation by exploiting function information from orthologous genes in another species. However, the information of orthologous genes is not always available for each gene of each species. Furthermore, this approach uses an unsupervised filtering approach (i.e., selecting function candidates with respect to the functions inferred from orthologous genes), while a supervised learning method for evidence combination may further improve the accuracy.

Xie et al. (2002) uses both textual information and sequence similarity scores for assigning GO terms to genes. Their automatic gene function prediction algorithm tries to identify unknown functions for genes and the focus of their work is not gene function

annotation in the literature. They use a clustering based approach for combining two types of evidence, while a more effective learning method can better adjust the weights of different types of evidence for more accurate results.

3 Methods

This section proposes a new framework (see Fig. 1) for gene function annotation in the literature by combining the three types of evidences as textual information, prior probability of gene functions and gene function information obtained from the known functions of genes with sequences similar to the target genes. We first present the method to extract useful features from the three types of evidence and then propose a supervised learning method to estimate the weights associated with the features.

3.1 Extracting useful features from multiple types of evidence

The task of gene function annotation in the literature is to annotate a specific target gene with corresponding Gene Ontology terms in a biomedical document (e.g., PubMed/MEDLINE abstract). For each biomedical document, the whole piece of text is segmented into separate sentences by a simple method that considers the period character. Furthermore, the stopwords (i.e., default Inquery stopword list included in Lemur[6]) and punctuation characters (i.e., any character other than an alphabetic letter or a digit) are removed from the text. The Porter stemmer[7] is further applied. The same procedures are applied on the GO term representations to remove stopwords and stem. For each sentence, a gene recognition algorithm (Bhalotia et al. 2003) is utilized to identify genes mentioned in the sentence by matching different variations of the gene names.

3.1.1 Features from textual information

Given a target gene mentioned in a specific sentence, we want to extract the evidence of available GO terms. The GO ontology used in this work (version of Dec 12, 2006) contains about 23,000 gene ontology terms. We extract the text from the name field and the exact synonym field as multiple text representations for each GO term. Therefore, a specific GO term G may have multiple text representations as $\left\{\overrightarrow{Tg}_1, \ldots, \overrightarrow{Tg}_K\right\}$, which contains K representations. Each text representation is a set of text tokens. All the tokens in each text representation of GO terms are stemmed using the Porter stemmer (e.g., the stemmed word representation for the GO term GO:0003824 "catalytic activity" is "catalyt active"). We assume that a single text representation $\overrightarrow{Tg}_k$ contains L tokens and is associated with a set of L stemmed tokens as $\overrightarrow{Tg}_k = \{w_{k1}, \ldots, w_{kL}\}$, the supporting evidence from the sentence *Sen* for this text representation is calculated as follows:

$$Ev(Sen, \overrightarrow{Tg}_k) = \sum_l (\delta(Sen, w_{kl}) * Weight(w_{kl})) \quad (1)$$

where the delta function δ is a binary function, which is 1 when the sentence contains this word and 0 otherwise. The Weight function indicates the importance of a specific word for

[6] http://www.lemurproject.org/

[7] http://snowball.tartarus.org/texts/introduction.html

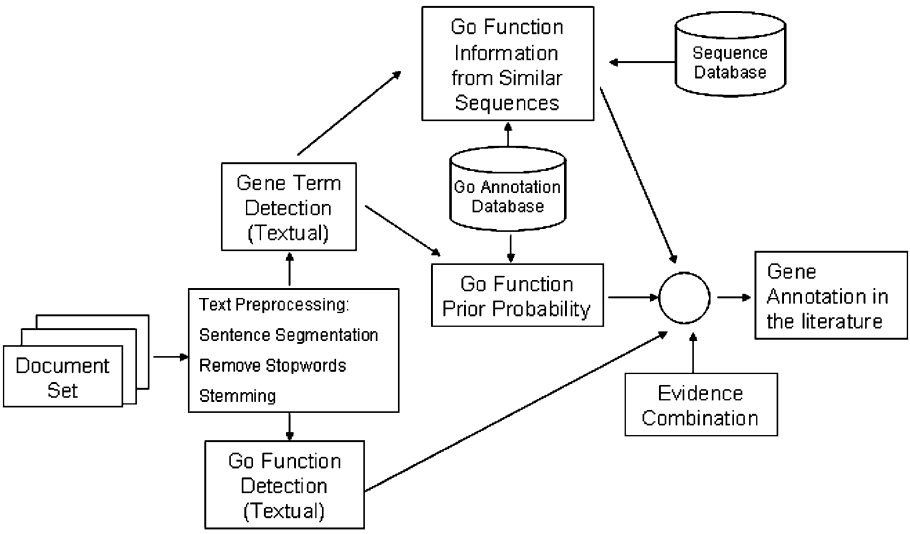

Fig. 1 The framework of combining gene sequence similarity and textual information for gene function annotation in the literature

gene function annotation. More specifically, the Weight function for a token is calculated as the inverse text representation frequency in a similar manner to the inverse document frequency in text information retrieval (Baeza-Yates 1999) as:

$$Weight(w_{kl}) = \log\left(\frac{N_{\text{Total\_textpresentation}}}{N_{\text{Total\_textpresentation}}(w_{kl}) + 0.5}\right) \tag{2}$$

where $N_{\text{Total\_textpresentation}}$ is the total number of text representations of GO terms and $N_{\text{Total\_textpresentation}}(w_{kl})$ is the total number of text representations of GO terms that contain this specific token w_{kl}. The intuition of the weight function is that a rare token in a text representation of a GO term is a more distinctive term than a much more common token that appears in many text representations of GO terms.

A longer text representation of a gene term generally has a larger value of the supporting evidence calculated by Eq. 1. In order to remove this bias, we calculated the normalized evidence with the value range between 0 and 1 in Eq. 3.

$$Ev'(Sen, \overrightarrow{Tg}_k) = \frac{\sum_l (\delta(Sen, w_{kl}) * Weight(w_{kl}))}{\sum_l Weight(w_{kl})} \tag{3}$$

Among all the text representations of a GO term G for a sentence, we choose a specific representation k^* that is associated with the largest value of the normalized evidence. Furthermore, among all the sentences of a document Doc, we choose the maximum supporting evidence from a specific sentence Sen^*.

Finally, given a target gene $Gene$, two features are defined to reflect the supporting evidence of the textual information in document Doc for a GO term G as:

$$f_{Text1}(Gene, Doc, G) = Ev'\left(Sen^*, \overrightarrow{Tg}_{k^*}\right)$$
$$f_{Text2}(Gene, Doc, G) = \sum_{l} (\delta(Sen^*, w_{k^*l}) * Weight(w_{k^*l})) \quad (4)$$

f_{Text1} is the normalized version of the text matching feature. f_{Text2} is introduced to consider the difference between a perfect match of a GO term (e.g., "reproduction") that is short and contains very common keyword(s) and a less perfect match of a GO term (e.g., "GPR37/endothelin B-like receptor activity") that is long and contains less common keyword(s). A less perfect match of a GO term that is long and contains less common keywords may be as indicative as a perfect match of a GO term that is short and contains very common keywords.

3.1.2 Features from the prior probability of gene function

When we consider annotating a target gene *Gene* with a specific GO term G in a biomedical document, it is generally useful to investigate how often this GO term is associated with other genes. More formally, we can calculate the prior probability that the specific GO term G is associated with an arbitrary gene as follows:

$$P_{prior}(G) = \frac{N_{gene}(G) + \beta}{N_{gene} + N_{gene} * \beta} \quad (5)$$

where N_{gene} is the total number of genes in a large curated database with many genes and the associated GO terms (i.e., Uniprot-SwissSprot database in this work); $N_{gene}(G)$ is the number of genes in the database that are associated with the specific GO term G. Since there are a large set of gene functions in the Gene Ontology, some gene functions may have not been instantiated in the biological database at all. To avoid assigning a zero prior probability to a GO term that does not occur in the biological database, the smoothing technique has been utilized for estimating the prior probabilities. Particularly, a pseudo count β is added for each gene product. The value of parameter β is set to 0.01 in this work. Therefore, the prior probability of a gene function that does not appear in the biological database in consideration becomes $\beta/(N_{gene}(1+\beta))$ instead of zero.

Finally, one feature of prior probability is defined with respect to the prior probability as:

$$f_{Prob}(Gene, Doc, G) = P_{prior}(G) \quad (6)$$

The Uniprot-SwissSprot database (release 51) is used in this work to calculate the prior probabilities.

3.1.3 Features of gene function derived from genes with sequences similar to a target gene

Features from textual information often have limited power to address the large variability in the expression of concepts in the biomedical literature. When a gene is annotated with a GO term in a piece of text, the tokens of the GO term may not occur together or only some of the tokens occur in the text. In contrast, even all the tokens of a GO term appear in a piece of text together with a gene, it may not be appropriate to annotate the gene with the GO term (e.g., the GO term is not directly associated the target gene).

Therefore, it is quite important to introduce another type of feature to boost the supporting evidence of GO terms correctly associated with a target gene and decrease the supporting evidence of GO terms incorrectly associated with the gene. To achieve this goal, we utilize the gene function information from the known functions of genes with sequences similar to the target gene. Our assumption is that genes with similar sequences often share similar functions.

Our approach to obtaining gene function information from genes with sequences similar to the target gene is similar to the PFP algorithm (Hawkins et al. 2006). Specifically, given a target gene *Gene*, conventional BLAST searches (version 2.2.15 blastall) (Altschul et al. 1990) are used to identify genes with sequences similar to the target gene (the default parameters were used). Please note that we do not use any annotations of sequences marked with the evidence codes IEA (Inferred from Electronic Annotation) and ISS (Inferred from Sequence Similarity) to avoid circular reference.

For a specific GO term G, the evidence that can be directly derived from the known functions of genes with sequences similar to the target gene *Gene* is calculated as follows:

$$Ev_{direct\_seq}(Gene, G) = \sum_j \big(\big(-\log(EValue(Seq_j)) + c\big) * I(G, Seq_j)\big) \tag{7}$$

where Seq_j is the jth top ranking similar sequence of the target gene. $EValue(Seq_j)$ denotes the corresponding E-value of the sequence similarity score obtained from BLAST. c is a constant that enables weakly similar sequences to be considered. We set c to be 2 in this work and thus allow the maximal E-value of similar sequence to be 100. Although we consider a large set of similar sequence by using a maximal E-value of 100, the logarithm weighting scheme enables most similar sequences (i.e., with small E-values) to have much higher weights than less similar sequences (i.e., with large E-values). I is an indicator function, which is 1 if the corresponding gene of Seq_j contains the function G and is 0 otherwise.

The above evidence is directly associated with the functions of genes with sequences similar to the target gene. Furthermore, it is reasonable to consider the supporting evidence for a GO term G from a known function G' of a gene having similar sequence to the target gene, if gene functions G' and G often co-occur but not identical. Therefore, the indirect evidence for GO term G can be defined as:

$$\begin{aligned} &Ev_{indirect\_seq}(Gene, G) \\ &\quad = \sum_j \big(\big(-\log(EValue(Seq_j)) + c\big) * S_{indirect}(G, Seq_j)\big) \end{aligned} \tag{8}$$

where $S_{indirect}(G, Seq_j)$ denotes the indirect score for function G from Seq_j, which is calculated by:

$$\begin{aligned} S_{indirect}(G, Seq_j) &= \max_{G' \in Seq_j \wedge G' \neq G} P(G|G') \\ P(G|G') &= \frac{N_{gene}(G, G') + \beta'}{N_{gene}(G') + N_{gene}(G') * \beta'} \end{aligned} \tag{9}$$

We calculate the indirect score by the gene function G' that has the largest conditional probability to be associated with the function G. N_{gene} is the total number of genes in a large curated database with many genes and the associated GO terms (i.e., Uniprot-SwissSprot database in this work); $N_{gene}(G')$ is the number of genes in the database that are associated with the specific GO term G', and $N_{gene}(G, G')$ is the number of genes in the

database that are associated with both the GO terms G and G'. Again, β' is set to 0.01 in this work.

Finally, given a target gene *Gene*, two features based on the gene function information from genes with sequences similar to the target gene are defined to reflect the direct and indirect evidence for the GO term G in a biomedical document *Doc* by calculating the normalized versions of the evidence as:

$$\begin{aligned} f_{direct\_seq}(Gene, Doc, G) &= \frac{Ev_{direct\_seq}(Gene, G)}{\sum_j \left(-\log(EValue(Seq_j)) + c\right)} \\ f_{indirect\_seq}(Gene, Doc, G) &= \frac{Ev_{indirect\_seq}(Gene, G)}{\sum_j \left(-\log(EValue(Seq_j)) + c\right)} \end{aligned} \tag{10}$$

3.2 Supervised learning method to combine multiple types of evidence

Section 3.1 introduces several useful features from multiple types of evidence for annotating genes with correct gene functions in the literature. It is an important task to combine these features together for accurate gene function annotation in the literature.

This paper proposes a supervised learning method to obtain a desirable model for combining the features. Particularly, given a set of features $\vec{f}$ associated with a target gene *Gene*, a GO term G and a biomedical document *Doc*, we use a logistic model to calculate the probability that the gene *Gene* is correctly annotated with the GO term G in the document *Doc* as:

$$P(correct(Gene, G, Doc)) = \frac{\exp\left(\sum_{f_i \in \vec{f}} \beta_i f_i + \beta_0\right)}{1 + \exp\left(\sum_{f_i \in \vec{f}} \beta_i f_i + \beta_0\right)} \tag{11}$$

where β_i is the weight associated with a specific f_i feature. β_0 is the weight associated with the bias constant feature (i.e., 1).

In particular, we study four types of models associated with three sets of features proposed in this work. The Comb_TI model only utilizes the two textual features as f_{Text1} and f_{Text2} as well as the bias const feature; the Comb_TI + Prior model utilizes the two textual features as f_{Text1} and f_{Text2}, the prior probability feature f_{Prob} and the bias feature; the Comb_TI + Seq model utilizes the two textual features as f_{Text1} and f_{Text2}, the two features calculated from sequence similarity scores as f_{direct_seq} and $f_{indrect\_seq}$, and the bias feature; the Comb_TI + Seq + Prior model utilizes all the features as the two textual features, the prior probability feature, the two sequence based features,, and the bias feature.

Given a small set of training data that includes some correct and incorrect examples of gene function annotations in the literature, we can estimate the desirable model parameters $\{\beta_i\}$ by maximizing the log likelihood of the training data. Particularly, the Quasi-Newton Method (Minka 2003) is used for estimating the model parameters.

4 Experimental methodology

Gene function annotation in biomedical literature has been studied within the BioCreAtIve (Blaschke 2005) and the TREC Genomics evaluation (Hersh et al. 2004; Seki and Mostafa 2005) tasks. In the BioCreAtIve task, participating systems are requested to assign exact GO terms to a target gene in biomedical documents. The TREC Genomics task is relatively easier, which requires participating systems to annotate a target gene with the three top level GO hierarchies (i.e., molecular function, biological process and cellular component) in biomedical documents.

Our purpose is to evaluate the accuracy of the proposed framework for assigning exact GO terms and to make comparison with the results in prior research. The BioCreAtIve task only provides 138 human genes and 99 biomedical articles for the test data. The results from the participating systems were inspected by human curators for the judgments. Therefore, it is difficult to obtain the judgments for the results generated by other systems after the evaluation task.

In this work, we utilize two large testbeds obtained from curated databases. More specifically, we use two testbeds as EBI Human and MGI Mouse (July 12, 2004 versions), where the results of gene function annotation by another two systems can be obtained from prior research (Stoica and Hearst 2006; Chiang and Yu 2003, 2004).

We only consider the GO annotations that are associated with PubMed abstracts. There are 13,626 GO terms associated with 4,410 human genes in 5,714 PubMed abstracts within the EBI human testbed. There are 1,947 mouse genes associated with 6,338 GO terms in the 2,188 PubMed abstracts within the MGI testbed. On the EBI testbed, about 85% GO terms are associated with 5 or less annotations. On the MGI testbed, about 89% GO terms are associated with 5 or less annotations.

Please note that curators manually make the GO annotations by looking through all the contents in the full-text publications, but we only have access to the PubMed abstracts of these publications. Since the information within the abstracts may not be sufficient to support the GO annotations, the performance of the evaluated algorithms is limited.

The Uniprot-SwissSprot database (release 51) is used to calculate the prior probabilities and conditional probabilities for GO terms. Before calculating the probabilities, all of the human and the mouse genes have been removed from the database.

BLAST searches are run against the Swiss-Prot gene sequence database to extract genes with sequences similar to a target gene. Particularly, all the human genes have been removed when we search for similar genes with respect to a target human gene; all the mouse genes have been removed when we search for similar genes with respect to a target mouse gene. This procedure is used in order to avoid the circular reference of existing knowledge in the curated databases so that the proposed algorithm does not take advantage of the direct function information within the databases.

Porter stemming has been used to preprocess the text data. Some other preprocessing steps are used to make the information within the two testbeds consistent with information from other databases.

The new proposed framework for gene function annotation in the biomedical literature uses a supervised learning approach, which requires a small amount of training data. We use cross validation for obtaining training and test data. For example, a 5-fold cross validation approach means that we first use the first one fifth data as the training data and the rest of the data as the test data, and then use the second one fifth data as the training data and the rest as the test data and so on. So generally, one fifth data is used as the training data. All the results from different splits are averaged to obtain the final results.

For each set of evidence, only a single model is created on each testbed for combining the evidence for all the GO codes, which can generate a robust model with a limited amount of training data. This single model is created to predict the probabilities of correct assignments of individual GO codes to target genes on every biomedical abstract in consideration. The single model works for all the GO codes, which may not exist in training data. Although not all GO codes appear in the training data, we can still estimate the overall contribution of different types of evidence from the training data. For a GO code that only appears in the test data, if it has a perfect match in the text (i.e., large values of textual features) and a high support from sequence based information (i.e., large values of sequence based features), the single model can combine the evidence and predict high probability of the correction assignment of this GO code.

Three evaluate metrics as precision, recall and *F*-measure score are used in this work. Precision denotes the percentage of correctly extracted gene function annotations among all the extracted gene function annotations. Recall denotes the percentage of correctly extracted gene function annotations among all the correct gene function annotations. The *F*-measure score (i.e., in short *F*-score) is the harmonic mean of precision and recall. In order to calculate precision, recall and the *F*-measure, we need to specify a threshold for the output probabilistic values of the evidence combination algorithm. In all the experiments, the threshold is automatically tuned on the training data. After the weights of the evidence combination method are obtained, the threshold is tuned to maximize the *F*-measure on the training data. The tuned threshold is further used on the test data.

5 Results and discussions

An extensive set of experiments were designed on the two testbeds to address the following questions of the proposed research:

(1) How good is the proposed framework by combining multiple types of evidences? Experiments are conducted to compare different versions of the proposed framework with different types of evidences as: (i) only the textual information (i.e., Comb_TI); (ii) textual information and the function prior probability (i.e., Comb_TI + Prior); (iii) textual information and gene function information from genes with similar sequences (i.e., Comb_TI + Seq); and (iv) all the three types of evidence of textual information, function prior probability and sequence based function information (i.e., Comb_TI + Seq + Prior).
(2) How good is the proposed framework compared with alternative solutions? We compare the results of the proposed framework with the results from prior solutions.
(3) How does the proposed framework work with different amount of training data? Experiments are conducted to evaluate the proposed framework when it is provided with different amount of training data for estimating the logistic regression model.

5.1 Experimental results by utilizing multiple types of evidence

The proposed framework in this work combines multiple types of evidence for gene function annotation in the biomedical literature. It is important to investigate the influence of different combinations of the evidence.

Table 1 Experimental results of the proposed framework by utilizing different types of evidences on two testbeds

Testbed	Feature set	Precision	Recall	*F*-score
EBI	Comb_TI	0.056	0.138	0.080
	Comb_TI + Prior	0.078	0.121	0.095 (+18.8%)
	Comb_TI + Seq	0.155	0.127	0.139 (+73.8%)
	Comb_TI + Seq + Prior	0.155	0.130	0.141 (+76.3%)
MGI	Comb_TI	0.053	0.155	0.079
	Comb_TI + Prior	0.111	0.136	0.119 (+50.6%)
	Comb_TI + Seq	0.180	0.149	0.161 (+103.8%)
	Comb_TI + Seq + Prior	0.177	0.152	0.162 (+105.6%)

The symbols of the feature sets include: TI: Text Information (2 features as described in Sect. 3.1.1); Prior: prior probability of gene function (1 feature as described in Sect. 3.1.2); and Seq: Evidence obtained from known functions of genes with sequences similar to the target genes (2 features as described in Sect. 3.1.3). All the results are obtained by using 5-fold cross validation, in which one fifth of the data is used as the training data and the rest of the data is used as the test data. For *F*-Scores, we use Comb_TI as the baseline results

The results of the four versions of framework are shown in Table 1. It can be seen from Table 1 that the Comb_TI method generates the least accurate results since it only utilizes simple textual features obtained by keyword matching. The Comb_TI + Prior method provides better results than Comb_TI, which shows that the prior information helps to improve the low baseline performance of using only textual information. Furthermore, the Comb_TI + Seq method and the Comb_TI + Seq + Prior method obtain substantially better results by incorporating more evidence. However, the performance of the Comb_TI + Seq method and the Comb_TI + Seq + Prior method are almost the same, which suggests that the prior information is not very helpful when the combination of textual information and sequence based information generates accurate results. This set of results clearly demonstrates the power of our framework to integrate multiple types of evidence for gene function annotation.

The advantage of incorporating the gene function information from genes with sequences similar to a target gene can be further demonstrated by the following examples.

The text of the first example is: "The localization of torsinA and torsinB immunoreactivity in neuronal processes points to a potential role for torsin proteins in synaptic functioning", which is from the abstract with the PubMed ID 11730696.

The Comb_TI method inappropriately suggests to annotate the gene torsinB (MOUSE) with the function "Protein Localization" (i.e., GO:0008104) because of the occurrences of the words "Localization" and "proteins".

Detailed analysis shows that none of the genes with sequences similar to torsinB has the function GO:0008104, which results in a zero value for the corresponding feature f_{direct_seq}. Furthermore, there is very limited indirect evidence from genes with sequences similar to torsinB, which results in a very small value for the corresponding feature $f_{indirect\_seq}$. Based on the new information, the Comb_TI + Seq + Prior method can successfully avoid the false annotation of torsinB with GO:0008104 by the Comb_TI method.

The text of the second example is: "Fusion of mouse CtBP1 or CtBP2 to Gal4DBD (Gal4 DNA binding domain) made them Gal4 binding site-dependent transcriptional repressors in transfected 10T1/2 cells, indicating their involvement in a transcriptional repression mechanism", which is from the abstract with the PubMed ID 10567582.

The Comb_TI method fails to annotate the target gene CtBP2 (MOUSE) with the function "transcription co repressor activity" (i.e. GO:0003714), since only two of the four keywords (i.e., transcriptional and repression) appear in the text.

On the other hand, the Comb_TI + Seq + Prior method finds that the top one gene (i.e., CtBP1_RAT) with sequence similar to CtBP2 (MOUSE) has the function GO:0003714. Furthermore, there is also enough indirect evidence from genes with sequences similar to CtBP2 (MOUSE). Finally, the Comb_TI + Seq + Prior method can successfully annotate the target gene CtBP2 (MOUSE) with the function GO:0003714 in the example.

5.2 Experimental results compared with results obtained from prior research

The section compares the results of Comb_TI + Seq and Comb_TI + Seq + Prior with the results from two prior solutions. The results of the Meke system are obtained from prior research in Chiang and Yu (2003). The Meke system uses more sophisticated textual information (e.g., phrase pattern matching) than the simple keyword matching information. The CSM + CSC system (Stoica and Hearst 2006) utilizes simple textual information and function information from orthologous genes in another species. The results of the Meke system and the CSM + CSC system are obtained from previous research in Stoica and Hearst (2006).

Table 2 shows the results of the four solutions. The results from Meke show high precision, but low recall and *F*-measure score. The *F*-measure scores of Meke system are still higher than the results of the Comb-TI method in Table 1 on both the two testbeds, which may be explained by the more sophisticated textual features used by the Meke system. The CSM + CSC system has consistently higher *F*-measure scores than the Meke system by utilizing the function information from orthologous genes. The Comb_TI + Seq and the Comb_TI + Seq + Prior methods achieve the highest *F*-scores on both testbeds. We believe that the advantage of our new framework comes from utilizing multiple types of evidence as well as the supervised learning method.

Table 2 Comparison of the experimental results of the proposed framework with the results obtained from prior research on these two testbeds

Testbed	Approaches	Precision	Recall	*F*-score
EBI	Meke	0.318	0.063	0.105
	CSM + CSC	0.163	0.092	0.118 (+12.4%)
	Comb_TI + Seq	0.155	0.127	0.139 (+32.4%)
	Comb_TI + Seq + Prior	0.155	0.130	0.141 (+34.3%)
MGI	Meke	0.332	0.049	0.086
	CSM + CSC	0.168	0.121	0.140 (+62.8%)
	Comb_TI + Seq	0.180	0.149	0.161 (+87.2%)
	Comb_TI + Seq + Prior	0.177	0.152	0.162 (+89.5%)

The results of the proposed framework (i.e., Comb_TI + Seq and Comb_TI + Seq + Prior) are generated by combining the three types of evidences as text information, prior probability of gene functions and gene function information obtained from known functions of genes with sequences similar to the target genes. For *F*-scores, we use Meke as the baseline results

5.3 Experimental results compared by utilizing different amount of training data

The proposed framework in this work utilizes a supervised learning method for combining multiple types of evidence. Therefore, it is important to investigate the effect on the accuracy by the size of training data.

Table 3 shows the results of the Comb_TI + Seq method with three training approaches, while Table 4 shows the results of the Comb_TI + Seq + Prior method with three training approaches. In the 5-fold cross validation approach, one fifth of all the data is used as the training data. In the 25-fold cross validation approach, one twenty-fifth of all the data is used as training data. In the 100-fold cross validation approach, only one hundredth of all the data is used as training data. It can be seen from both Tables 3 and 4 that the two training approaches of 5-fold and 25-fold cross validation generate competitive results, while the approach of 100-fold cross validation obtains a bit worse results. However, the results of 100-fold configuration are still better than or at least as good as the results from previous systems (i.e., CSM + CSC and Meke from Table 2). This indicates that the new framework works reasonably well even with a limited amount of training data. This can be explained by the fact that only a small number of model parameters are used by the

Table 3 Experimental results of the proposed framework with the Comb_TI + Seq method when different amount of training data is available

Testbed	Training configuration	Precision	Recall	*F*-score
EBI	5-fold	0.155	0.127	0.139
	25-fold	0.156	0.126	0.137
	100-fold	0.141	0.131	0.128
MGI	5-fold	0.180	0.149	0.161
	25-fold	0.170	0.157	0.160
	100-fold	0.138	0.166	0.142

The results of the proposed framework are generated by combining all the three types of evidences as described in Table 1. The results are obtained in three configurations: one is 5-fold cross validation where one fifth of the data is used as training data, one is 25-fold cross validation where one twenty-fifth of the data is used as training data, and another is 100-fold cross validation where one hundredth of the data is used as training data

Table 4 Experimental results of the proposed framework with the Comb_TI + Seq + Prior method when different amount of training data is available

Testbed	Training configuration	Precision	Recall	*F*-score
EBI	5-fold	0.155	0.130	0.141
	25-fold	0.157	0.130	0.139
	100-fold	0.126	0.146	0.129
MGI	5-fold	0.177	0.152	0.162
	25-fold	0.173	0.158	0.160
	100-fold	0.135	0.174	0.141

The results of the proposed framework are generated by combining all the three types of evidences as described in Table 1. The results are obtained in three configurations: one is 5-fold cross validation where one fifth of the data is used as training data, one is 25-fold cross validation where one twenty-fifth of the data is used as training data, and another is 100-fold cross validation where one hundredth of the data is used as training data

Comb_TI + Seq and the Comb_TI + Seq + Prior methods, which can be estimated by a limited amount of training data.

6 Conclusions

Gene function annotation in the biomedical literature is an important research problem. However, most current solutions only utilize textual information and their performance is not satisfactory due to the large variability in the expression of concepts in biomedical literature.

This paper proposes a framework for gene function prediction in the biomedical literature by combining multiple types of evidence as: textual information, the prior probabilities of gene functions and the gene function information from the known functions of genes with sequences similar to a target gene. A supervised learning method is utilized to obtain the weights for combining the three types of evidence together for gene function annotation. An extensive set of experiments on two testbeds has shown the benefits of combining multiple types of evidence (particularly the sequence-based information) and the advantage of the proposed framework against two prior solutions.

There are several directions to improve the research in this work. More sophisticated text features (e.g., from text classification techniques) can be incorporated into the framework for more accurate results. It is also promising to explore more advanced learning methods for combining the multiple types of evidence.

References

Altschul, S. F., et al. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*, 403–410.

Baeza-Yates, R. (1999). *Modern information retrieval*. New York: ACM Press.

Bhalotia, G., et al. (2003). Biotext team report for the TREC 2003 genomic track. In *Proceedings of TREC 2003*.

Blaschke, C. (2005). Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, *6 Suppl 1*, S16.

Camon, E., et al. (2004). The gene ontology annotation (goa) database – an integrated resource of go annotations to the uniprot knowledge base. *In Silico Biology, 4*(1), 5–6.

Chiang, J. H., & Yu, H. C. (2003). Meke: Discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics, 19*(11), 1417–1422.

Chiang, J. H., & Yu, H. C. (2004). Extracting functional annotations of proteins based on hybrid text mining approaches. In *Proceedings of BioCreative Workshop*.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6*, 57–71.

Couto, F., Silva, M., & Coutinho, P. (2005). Finding genomic ontology terms in unstructured text. *BMC Bioinformatics*, *6 Suppl 1*, S21.

Ehrler, F., & Ruch, P. (2005). Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot. *BMC Bioinformatics, 6 Suppl 1*, S23.

Eisen, M. B., et al. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America, 95*(25), 14863–14868.

Hawkins, T., et al. (2006). Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Science, 15*, 1550–1556.

Hawkins, T., & Kihara, D. (2007). Function prediction of uncharacterized proteins. *Journal of Bioinformatics and Computational Biology, 5*, 1–30.

Hersh, W. R., et al. (2004). TREC 2004 genomics track overview. In *Proceedings of TREC 2004*.

Jensen, L. J., et al. (2003). Prediction of human protein function according to gene ontology categories. *Bioinformatics, 19*(5), 635–642.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, 21–23 April (pp. 137–142).

Joslyn, C. A., et al. (2004). The gene ontology categorizer. *Bioinformatics, 4*(20), 1169–1177.
Kim, W., & Wilbur, W. J. (2005). A strategy for assigning new concepts in the MEDLINE database. In *Proceedings of AMIA Symposium*, 2005.
Koike, A., Niwa, Y., & Takagi, T. (2004). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics, 21*(7), 1227–1236.
Krallinger, M., Padron, M., & Valencia, A. (2005). A sentence sliding window approach to extract protein annotations from biomedical articles. *BMC Bioinformatics, 6 Suppl 1*, S19.
Marcotte, E. M., et al. (2000). Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America, 97*(22), 12115–12120.
Minka, T. (2003). *A comparison of numerical optimizers for logistic regression*. Unpublished draft.
Ray, S., & Craven, M. (2005). Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics, 6 Suppl 1*, S18.
Raychaudhuri, S., et al. (2002). Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research, 12*(1), 203–214.
Rice, S. B., et al. (2005). Mining protein function from text using term-based support vector machines. *BMC Bioinformatics, 6 Suppl 1*, S22.
Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.
Ruch, P., Perret, L., & Savoy, J. (2005). Features combination for extracting gene functions from MEDLINE. In *Proceedings of European Colloquium on Information Retrieval (ECIR)*.
Sebastiani, F. (1999). *Machine learning in automated text categorisation*. Paris, France: Centre National de la Recherche Scientifique.
Seki, K., & Mostafa, J. (2005). An application of text categorization methods to gene ontology annotation. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
Settles, B., & Craven, M. (2004). Exploiting zone information, syntactic features, and informative terms in gene ontology annotation from biomedical documents. In *Proceedings of TREC 2004*.
Stoica, E., & Hearst, M. (2006). Predicting gene functions from text using a cross-species approach. In *Proceedings of Pacific Biocomputing Symposium*.
Verspoor, K., et al. (2004). Protein annotation as term categorization in the gene ontology. In *Proceedings of BioCreative Workshop*.
Wu, C. H., et al. (2006). The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Research, 34*(Database Issue), D187–D191.
Xie, H., et al. (2002). Large-scale protein annotation through gene ontology. *Genome Research, 12*(5), 785–794.
Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*, 69–90.