Literature Based Functional Annotation of Genes

Pooja Jain[1*], Francisco M. Couto[1], Mário J. Silva[1], Jörg D. Becker[2]

[1]Departamento de Informática

Faculdade de Ciências, Universidade de Lisboa, Portugal

[2]Instituto Gulbenkian de Ciência, Oeiras, Portugal

[*]pjain@xldb.di.fc.ul.pt

Abstract

This paper proposes a fast and automatic functional annotation method for hundreds of unannotated genes in biological databases. Genes are often described and discussed in biomedical literature that can be used to unravel the functions of genes. We have proposed a reliable text mining approach to discover functional annotations for genes from literature. This approach was validated by building the APEG (Arabidopsis Pollen Expressed Genes) database system, which integrates information about 147 pollen selectively expressed genes of *Arabidopsis thaliana*. APEG operates with ProFAL, a text mining and automatic database annotation tool. The automatically extracted annotations for genes were evaluated by comparing them with those obtained by curating the same literature. Functional annotations were extracted with an average precision and recall of 61% and 78%, respectively with identification of 21 novel probable functions for 8 genes. The results show that mining the biomedical literature can effectively increase our knowledge about genes.

Key words: Text mining, Functional Annotation, Biological Databases, Arabidopsis thaliana, Pollen.

1 Introduction

The current applications of information extraction and text mining to molecular biology are becoming increasingly important because of the growth of the related data and scientific knowledge. To cope with this growth, databases and the scientific literature have to be integrated and information therein has to be filtered and categorized. The work presented in this work addresses this need.

A large number of curators are engaged with manual assignment of functional annotations to genes using various methods. One of these methods is reading and extracting important knowledge from biomedical literature. However, this manual method is unable to keep pace with the rate of accumulation of data and may be affected by the restricted domain knowledge of the individual performing the task. We present a fast, robust and automatic context-specific approach for analysis of literature data that is heterogeneous in structure, contents and semantics. The approach uses a text mining tool to automatically extract functional annotations in form of GO terms for a collection of genes organized in a relational database system.

A *functional annotation* represents an association between a gene and a GO term describing its function in the biological database. This combination gives, for each identified functional annotation, an evidence in the form of the sentence identified in the published literature. Curators always look for this kind of evidence while curating a biological entity with great reliability and precision. Moreover, assigning GO terms as functional annotations increases the annotations coverage for all possible functions, an observation also reported by others [10]. Rison et al. have also described GO as *representative of the 'next generation' of functional schemes* [13].

The problem of functional annotation is already been addressed using text mining techniques. In the BioCreAtIvE workshop (Critical Assessment of Information Extraction systems in Biology) the GO based, automatic extraction of functional annotations to proteins from full-text articles resulted in a perfect prediction percentage equal to 11.80% [2]. Similarly in GOA (Gene Ontology Annotation) database, the GO terms identified electronically as well as manually are assigned as functional annotations to biological entities [4]. The enzymes in CAZy (Carbohydrate Active enZymes), a database of carbohydrate active enzymes, were annotated with a text mining tool ProFAL (bioPROducttein Functional Annotation through Literature) [6], [7].

2 APEG

APEG (Arabidopsis Pollen Expressed Gene) is a relational database system for managing information about pollen selectively expressed genes of *Arabidopsis thaliana* [11]. Currently, APEG maintains information about 147 pollen selectively expressed genes identified in a separate study on the Arabidopsis pollen transcriptome [9]. The information about genes include their cross references to public databases, expression analysis results and their functional annotations, automatically extracted from the literature. It also links genes to TAIR, GenBank and SwissProt entries and to the relevant publications available at PubMed.

APEG has a web based interface that provides an intuitive and user friendly way to load data from external sources, mine that data, and perform the final verification steps of our approach. Its main features are:

- Search and query APEG's data.
- Access cross-references to other database entries.
- Access bibliographic references to the genes.
- Access automatically extracted functional annotations to the genes.
- Provide curator specific features to manually add the functional annotations and bibliographic references to the genes and to mark the extracted, valid functional annotations to the genes.

APEG is available on line at : `http://xldb.fc.ul.pt/rebil/tools/apeg/` .

3 Our Approach

We have developed a relational database system APEG to organize gene expression data [11]. Once organized in a database, the gene information can be augmented with knowledge from other databases or the literature, and may be accessed for browsing

and curation by domain experts. Cross-references for genes to public databases, such as GenBank [3], SwissProt [1] and TAIR (The Arabidopsis Information Resource) [5] were collected. These cross-references were used to obtain the bibliographic references to the articles relevant to the genes. We applied ProFAL [6] to retrieve freely available abstracts of these articles from PubMed [12] and to extract functional annotations for genes, using GO terms from GOA project. ProFAL validates each of the extracted annotations using a quantitative measure based on the information content of each word in the term and assigns a confidence score to them [8]. Lastly, we compared automatically extracted annotations with the manually extracted annotations through the database user interface.

We have validated our approach by applying it to a collection of pollen selectively expressed genes of *Arabidopsis thaliana*, identified in a comparative expression study [9]. The approach identified multiple functional assignments, some of them were already known, whereas some others are regarded as probable novel functions.

3.1 Annotation Curation and Verification

We evaluated our approach by checking the results of the automatic extraction by ProFAL. Extraction was performed with different values of Extraction Parameter (EP) from 0.6 to 1.0. The higher the value of EP, the more stringent will be the extraction process. The annotations extracted with different values of EP, were curated separately by a human curator through the APEG user interface and verified as:

1. True positives - The extracted annotations, which were *accepted as valid* by the curator. These were also extracted manually and therefore curated as correct functional annotations

2. False positives - The extracted annotations, which were *not accepted as valid* functional annotations.

3. False negatives - The annotations, which were curated manually for a gene but were not extracted by ProFAL.

4. Probable functional annotations - The extracted annotations, which were not extracted manually, but *accepted as valid* by the curator and judged as a possible function.

The total number of true positives, false positive and false negatives were counted to calculate precision and recall for every gene. Finally, the average values of precision and recall for every run of the ProFAL were calculated.

4 Results and Discussion

Annotations were extracted from the abstracts of a total of 55 distinct articles for 48% of the genes. For detailed document retrieval statistics please see [11]. The annotations verification statistics from five different extraction cycles are summarized in Table 1. 'GO Terms' represents the total number of GO terms used to derive the annotations for the genes in the corresponding extraction run. The detailed results of verification are available from the APEG site at: `http://xldb.fc.ul.pt/rebil/tools/apeg/annot_veri.php`.

The automatic gene functional annotation approach as presented in this paper is capable of extracting functional information latent in biomedical literature without significant manual effort on downloading and reading documents relevant to the genes.

Table 1: Annotation verification statistics from five different annotation extraction runs using different values of the Extraction Parameter (EP)

Extraction Cycles	EP=0.6	EP=0.7	EP=0.8	EP=0.9	EP=1.0
False Positives	628	138	135	86	81
True Positives	60	61	16	31	9
False Negatives	59	69	89	105	134
Probable Functions	16	27	9	10	7
GO Terms	250	173	151	148	125
Precision	33.31	61	59.67	76.39	62.86
Recall	72.51	77.63	64.91	55.08	63.81

Table 2: Annotations and respective evidence text from the article.

Annotation	**Evidence text**
protein homooligomerization activity	By analysing C-terminal deletion forms of Hsp17 class II, we obtained evidence that the intact C-terminus is critical for the oligomerization state, for the heat-stress-induced auto-aggregation and for recruitment of class I proteins
chaperone activity	Heat-stress granules (HSG) are highly ordered, cytoplasmic chaperone complexes found in all heat-stressed plant cells.

Retrieving relevant articles to be mined, by using bibliographic references from public databases was effective in detecting the relevant literature. Only 6 articles out of 55 distinct articles were found to be too general.

We determined the optimum value for EP as 0.7, based on the average values of precision and recall for the respective extraction cycle (Table 1). The average precision and recall values calculated for this run were 61% and 77.63% respectively. While determining the optimum value for EP, a high recall is preferred over a high precision, to avoid loss of any possible functional annotation for a gene. Although there is always a trade off between including every possible functional annotation and the time a human reader has to invest on selecting the true annotation. As expected, for higher values of EP, we have less GO terms with a larger confidence on being correctly identified.

4.1 Specific Annotations

The verification results confirms that our approach is superior to the manual curation and is capable of extracting more specific annotations. This is due to the consideration of all the graphs in GO and use of deeper GO terms than using the more general terms higher in the GO hierarchy often considered in manual curation. For example, the first of the extracted annotations in Table 2 "cytoskeleton organization and biogenesis" is more specialized term to describe the cytoskeleton related processes governed by molecular motors in a cell.

4.2 Probable Functions

The proposed approach was also able to identify 21 probable functional annotations for 8 gene out of 64 genes for which annotations were extracted. These annotations corresponds to the *accepted valid annotations* but were not stated as such in the literature. For details

please see [11]. These results could give guidelines to confirm the probable functions experimentally.

5 Conclusions

We have evaluated the performance of a text mining tool towards automatic functional annotation of genes contained in a database system APEG. The annotations were extracted with a precision and recall of 61% and 78% respectively. This work has concluded that:

1. Text mining is effective in extracting tacit and specific knowledge pertaining gene's function.
2. The extraction of specific evidence texts indicating the function of a gene from different documents reduces the time and need to read the entire document to identify the functions it describes.
3. Extracts automatically the functional annotations to genes, which were already assigned to them by human experts. This demonstrates its potential as an assistant for database curators, performing automatic database annotation.

Some future improvements were also identified, such as retrieving relevant articles for all the genes improving percentage of genes for which annotations could be extracted. Efforts to decrease the false positive annotations and extension of the validation work with a more comprehensive set of genes, or a direct comparison with other systems could be some other interesting future directions.

References

[1] B. Boeckmann et. al. *Nucl. Acids. Res.*, 31(1):365–370, 2003.

[2] BioCreAtIvE. `http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html`, 2003.

[3] D. A.Benson et. al. *Nucl. Acids. Res.*, 28(1):15–18, 2000.

[4] E. Camon et. al. *Nucleic Acids Res.*, 32:262–266, 2004.

[5] E. Huala et.al. *Nucl. Acids Res.*, 29(1):102–5, 2001.

[6] F. Couto et. al. In *VIII Conference on Software Engineering and Databases (JISBD)*, pages 747–756, 2003.

[7] F. Couto et. al. In *Data Mining and Text Mining for Bioinformatics (PKDD)*, 2003.

[8] F. Couto et. al. In *BioCreAtIvE*, Granada, Spain, March 2004.

[9] J. D. Becker et. al. *Plant Physiology*, 133:713–725, 2003.

[10] J. Schug et. al. *Genome Research*, 12:648–655, 2002.

[11] P. Jain. DI/FCUL TR –, Department of Informatics, University of Lisbon, 2004.

[12] PubMed. `http://www4.ncbi.nlm.nih.gov/`.

[13] S. Rison et. al. *Functional Integrative Genomics*, 1:56–69, 2000.