Mining the Biomedical Literature for Genic Information

Catalina O. Tudor ⋆ **K. Vijay-Shanker** ⋆ **Carl J. Schmidt** ◦
Department of Computer and Information Sciences ⋆
Department of Animal and Food Sciences ◦
University of Delaware, Newark, DE 19716
{tudor,vijay}@cis.udel.edu schmidtc@udel.edu

Abstract

eGIFT (**E**xtracting **G**ene **I**nformation **F**rom **T**ext) is an intelligent system which is intended to aid scientists in surveying literature relevant to genes of interest. From a gene specific set of abstracts retrieved from PubMed, eGIFT determines the most important terms associated with the given gene. Annotators using eGIFT can quickly find articles describing gene functions and individuals scientists surveying the results of high-throughput experiments can quickly extract information important to their hits.

1 Introduction

Given the huge number of articles from the biomedical domain, it has become very difficult for scientists to quickly search and find the information they need. Systems to facilitate literature search are being built. E.g. GoPubMed (Doms and Schroeder, 2005) clusters abstracts retrieved from PubMed based on GO and MeSH terms, iHOP (Hoffman and Valencia, 2005) connects biomedical literature based on genes, EBIMed (Rebholz-Schuhmann et al., 2006) displays sentences containing GO terms, drugs, and species.

In contrast to these systems, eGIFT automatically identifies the most relevant terms associated with a given gene. We believe that such a retrieval of terms could itself enable the scientists to form a reasonable good idea about the gene. For example, some of the top key phrases associated with *Groucho* (Entrez Gene ID 43162) by eGIFT are: *transcriptional corepressor*, *segmentation*, *neurogenesis* and *wd40*. This might immediately inform a user that *Groucho* is probably a *transcriptional corepressor*, that it might be involved in the processes of *segmentation* and *neurogenesis* and that it might contain the *wd40* domain, which allows them to draw further inferences about the gene. To enable the scientists to get a deeper understanding, eGIFT further allows the retrieval of all sentences from this gene's literature containing the key phrase in question. The sentences can be displayed in isolation or in the context of the abstract in which they appear.

2 Ranking Key Terms

(Andrade and Valencia, 1998) automatically extracted keywords from scientific text by computing scores for each word in a given protein family, based on the frequency of the word in the family, the average frequency of the word and the deviation of word distribution over all families. (Liu et al., 2004) extended this method to statistically mine functional keywords associated with genes.

Our application is somewhat similar in that we compare the distribution of phrases in the abstracts about the gene from some background set. We use statistical methods to identify the situations where the different frequencies of appearance of a term in two sets of the literature are statistically interesting. We differ from the above work by choosing a broader range of background information. Our motivation is to retrieve any type of phrases, thus not limiting ourselves to only functional terms or terms that might differentiate the selected set of protein families. Since we no longer have several sets of litera-

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 28–29,
Columbus, Ohio, USA, June 2008.

ture, our approach differs from the above method in that we cannot base the score on average frequencies and term deviation in the same way.

Background Set (BSet): In order to capture a wide range of information about genes in general, we downloaded from PubMed all the abstracts for the following boolean query: gene[tiab] OR genes[tiab] OR protein[tiab] OR proteins[tiab]. Approximately 640,000 non-empty abstracts were found.

Query Set (QSet): We download from PubMed the abstracts that mention a given gene name and its synonyms. We obtained the latter from BioThesaurus (Liu et al., 2005).

Key Term Scores: We considered many different statistical tests to identify significant key phrases, but eventually settled on the following score:

$$s_t = (\frac{dc_{tq}}{N_q} - \frac{dc_{tb}}{N_b}) * ln\left(\frac{N_b}{dc_{tb}}\right)$$

where dc_{tb} and dc_{tq} are the background and query document counts of term t, and N_b and N_q are the total number of documents from the BSet and QSet.

The difference in frequencies ($\frac{dc_{tq}}{N_q} - \frac{dc_{tb}}{N_b}$) gives preference to terms that appear more frequently in the QSet than in the BSet. This way, we would like to capture terms that are common to the given gene but not to genes and proteins in general. The difference itself is not sufficient to eliminate common words. To address this problem, similar to the use of IDF in IR, we add a global frequency term ($ln\left(\frac{N_b}{dc_{tb}}\right)$) to further penalize common terms, such as *protein*.

To better understand how the score is computed, consider the gene *Groucho* and its key term *corepressor*, which was mentioned in 66% of the QSet and only in 0.1% of the BSet. The huge difference in frequencies, together with the low background frequency, helped the key term *corepressor* score 4.3617, while most of the terms score below 0.25.

Enhancements to Basic Method: First, we extended our method to include unigrams, bigrams, and multi-word terms where previously identified. We observed that some words are not meaningful when presented alone. For instance, the words *development* and *embryonic* taken separately are not as informative as when put together into *embryonic development*, a term which was ranked much higher than the two words.

Next, we applied morphological grouping on terms, based on manually developed rules, after observing variances within the same concept. In writing, we can say *corepressor*, *co-repressor*, or *co-repressors*. In order to capture the concept, we computed frequencies on morphological groups and not on each individual term.

Last, we divided key terms into categories by using morphological information to separate terms such as descriptors, and by consulting publicly available controlled vocabularies (such as NCBI Conserved Domains, NCBI Taxonomy, MedlinePlus, DrugBank, and MeSH category A01).

3 Assessment

Our method has been applied on 55 different genes selected by annotators for a public resource. The initial feedback has been encouraging. Also preliminary investigations suggest we get far more keywords associated with some genes in resources such as GenBank, SwissProt and Gene Ontology than the system of (Liu et al., 2004). Our next goal is to do a thorough evaluation of our system.

References

Miguel A Andrade and Alfonso Valencia. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–607.

Andreas Doms and Michael Schroeder. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleid Acid Research*, 33:w783–w786.

Robert Hoffman and Alfonso Valencia. 2005. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21:ii252–ii258.

Ying Liu, Martin Brandon, Shamkant Navathe, Ray Dingledine, and Brian J. Ciliax. 2004. Text mining functional keywords associated with genes. *MedInfo*, 11:292–296.

Hongfang Liu, Zhang-Zhiu Hu, Jian Zhang, and Cathy Wu. 2005. Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22(1):103–105.

Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui, Sylvain Gaudan, Mark Riethoven, and Peter Stoehr. 2006. EBIMed - text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23:e237–e244.