Mining for Gene-Related Key Terms: Where Do We Find Them?

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

Abstract

This paper is concerned about one aspect in the extraction of key terms that describe various types of information about a given gene. Our method for key term extraction is based on a comparison of term occurrences in documents associated with the gene versus a broader set of documents. We investigate the influence on the type of key terms extracted by the type of documents retrieved for the given gene. We provide analysis on five genes to draw our conclusions and hypotheses for future investigations.

1 Introduction

Researchers spend a tremendous amount of time searching the biomedical literature for information they need. A simple PubMed query for a specific gene can sometimes return several thousands of articles, which could be time consuming to read. Instead, we allow researchers to consult a list of most important gene-related information (key terms) gathered automatically from these articles. By consulting key terms and by reading sentences containing a particular key term, the researchers can find quickly information of interest.

For example, searching PubMed for abstracts containing gene *Groucho* returns a list of 269 references to articles. We identify key terms and present users with relevant information: *transcriptional corepressor*, *segmentation*, *neurogenesis* and *WD40*. This immediately informs a user that *Groucho* is a *transcriptional corepressor*, that it might be involved in the processes of *segmentation* and *neurogenesis* and that it might contain the *WD40* domain. From these key terms, researchers can choose to learn more by reading sentences and abstracts containing the terms of interest.

We determine such key terms by comparing the set of documents retrieved for the specific gene (the query set) against a background set of documents with information about genes in general. The type of documents retrieved may influence the type of information captured by the extracted key terms. We investigate how different kinds of key terms can be obtained based on changing the query set. We report our findings about the type of key terms we extracted for five genes when using different query sets. We believe these findings about the influence of the different query sets are not limited to our method for key term extraction, but also to all key term extraction systems that consider term distributions between a background set and a set associated with a given gene.

2 Related Work

One of the earliest works on mining key terms from text is due to Andrade and Valencia (1998). They proposed to automatically mine keywords for families of proteins, by comparing each family's literature against the other families' combined literature. Other systems which also mine key terms from the biomedical literature are built: *e-LiSe* (Gladki et al., 2008), *MedEvi* (Kim et al., 2008), and *Anne O'Tate* (Smalheiser et al., 2008). Our system, eGIFT, **E**xtracting **G**ene **I**nformation **F**rom **T**ext (Tudor et al., 2008), differs from these systems in its intended use only for genes; the construction of background information; the filtering of irrelevant documents; the extension of words to multi-word key terms; the grouping of morphologically related terms; and the division of key terms into categories.

3 Retrieving key terms using eGIFT

We compare the distribution of terms in the abstracts about the gene from some background set. We look for situations where the different frequencies of appearance of a term in two sets of the literature are statistically interesting. For the **Background Set**, we downloaded from PubMed all abstracts for the search on *gene(s)* or *protein(s)*. For the gene-specific documents, we download abstracts from PubMed which mention a given gene name and its synonyms, and call it the **Query Set**. Using these sets of documents we compute the score s_t for a term t as follows:

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

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

The difference between the normalized document frequencies ($\frac{dc_{tq}}{N_q} - \frac{dc_{tb}}{N_b}$) is giving preference to terms that appear more frequently in the Query Set than in the Background Set, while the second part of the equation ($ln\left(\frac{N_b}{dc_{tb}}\right)$) further penalizes common terms in general. We rank the key terms based on their scores, in decreasing order.

4 Research Methods

We have applied our method on 60 genes selected by annotators for a public resource. A set of 5 genes was chosen for our analysis by one of the co-authors expert in Biology and familiar with the selected genes. Their symbols and Entrez Gene IDs are: BMP2 650, GRO 43162, LMO2 4005, OPN 6696, and TERT 7015. Together, we determined the category of each key term, and for each gene we compared the results returned by the different query sets, as described below. For each set, we looked at the top 150 key terms only.

Since the primary goal of this work is to determine how the choice of gene-specific set of documents influences the quality and type of information extracted, we consider for a given gene many different query sets, as will be defined next.

We observed that not all the abstracts from the Query Set are relevant to the given gene. When we search for a specific gene, we obtain two types of abstracts: (1) which talk mainly about the gene, and (2) which are focused primarily on some other topic but happen to mention our gene. Given this observation, we have decided to divide the entire set of retrieved documents for a gene (**Full Set**) into two distinct sets: **About Set** and **Extra Set**. By considering the About Set, instead of the Full Set, we hope to filter out information which is not core to the given gene. We check if an abstract mentions the given gene at least three times, or once in the title, the first or last sentence of the abstract, before assigning it to About Set.

While we expect to obtain more "core" key terms by using About Set as the query set, we also want to see what kind of key terms are found when we use Extra as the query set. However, since Extra documents are supposed to be about some other topic and might just mention our gene, we can focus on the sentences, in the Extra abstracts, that contain our gene, as this might give us gene-related information when mentioned in context of some other topic. So we build a new possible query set, **ExtraSent Set**, that is obtained by taking each document in the Extra Set and only retaining sentences that mention our gene. We similarly obtain **AboutSent** and **FullSent** sets.

Since the title, first and last sentences of the abstracts generally give a high level summary of the work they discuss, we create **AboutTiFL** by only retaining the title, first and last sentences. By using AboutTiFL as the query set, we expect to do well on extraction of high level key terms, but not more detail level key terms [1] , for the gene.

5 Discussion of Results

5.1 About Set vs. Full/Extra Set

As we expected, the use of About as the query set led to better extraction of information that is core to the given gene. For example, processes like *segmentation*, *neurogenesis*, *embryonic development*, and *sex determination* are ranked much higher in the About Set than in the Extra Set for gene *Groucho*. *Groucho* is involved in all of these processes, and since many abstracts "about *Groucho*" will discuss its functions and processes, these terms are highly ranked in contrast to the use of Full Set or Extra Set as the query set. Since the Extra Set abstracts aren't necessarily about *Groucho*, these key terms are ranked much lower and some other key terms take their place in the Extra Set ranking. We found that the highly ranked key terms for the Full Set include terms

[1] By high level we mean process/functional terms, and by detail level terms we mean other genes and domains/motifs

from both About and Extra and the four processes drop in rank, particularly *embryonic development* and *sex determination*. We see several such cases. For example, consider the association of *Lmo2* with *erythropoiesis*. *Lmo2* was originally identified as an oncogenic protein in human t-cell leukemia and later determined to be essential for erythropoiesis (PMID 9520463). *Chromosomal translocations*, *erythropoiesis*, *tumorigenesis*, and *t-cell development* are ranked higher in About than in Full, and, in fact, with the Extra Set the rank dropped considerably. For the gene *Opn*, *secretion*, *cell adhesion*, and *metastasis* ranked very high in the About Set, while only one of these terms ranked in the top 150 key terms for Extra Set.

In contrast, the use of Extra Set as the query set reveals some highly interesting and potentially useful information about the genes which get ranked much lower in the About Set. Rather than high level process/function oriented key terms, with Extra ranking we are able to extract information that is often "lower level", such as other related genes and domains/motifs. Although some of the key terms obtained by using Extra Set are relevant to the given gene, many are "false positives" (i.e. highly ranked terms that were not associated with the gene).

5.2 Sentence-based Document Sets

ExtraSent Set. Extra Set contains many terms that are extraneous to our gene. Hence, we propose to investigate the use of ExtraSent Set as this might filter out terms less relevant to our gene. We notice that this is exactly the situation. Genes and motifs retrieved by using the Extra Set get ranked even better with ExtraSent Set. For example, *eh1* and *bhlh*, which are highly ranked in ExtraSent as compared to About, are domains that are contained in other genes which interact with *Groucho*. Abstracts that focus on other topics/genes but which also mention *Groucho* (and hence make it into Extra Set of *Groucho*) discuss *eh1* and *bhlh* frequently.

Also, some genes are highly ranked with ExtraSent Set when they co-occur frequently with our gene. This might happen when several genes are mentioned together because they form a complex, participate in some pathway, contain a common motif, are expressed in some disease, etc. For example, the gene *Lyl1*, is mentioned by Entrez Gene for interacting with *Lmo2*. ExtraSent is the only set which includes *Lyl1* in the first 150 key terms and ranks it at the top of its list.

Another example is *activin* to be discussed in the context of *Bmp2*. *Activin* is in many ways similar to *Bmp2*, and somebody interested in *Bmp2* would want to know this information. But in particular we believe that the relevance of *activin* can be noted in that some sentences not only discuss similarities, but go on to point out some small but significant differences: "... human CHL2 (hCHL2) protein is secreted and binds activin A, but not BMP-2 ..." (PMID 15094188) and "... BMP-2 and activin A induce PC12 cell neuron differentiation ..." (PMID 8663261). So in some sense, *activin*, while not central to *Bmp2*, may be important to researchers interested in *Bmp2*. *Activin* does not rank highly in the About Set (rank 157), nor in FullSent Set (rank 106), but gets a much higher rank of 25 in ExtraSent Set (while in Extra Set it has rank 90).

A similar example can be noticed with the gene *Opn*. Two genes were boosted in the ExtraSent Set (*DMP-1* and *DSPP*) which were otherwise not present in any of the top 150 key terms for the other sets. *Opn*, *DMP-1* and *DSPP* are SIBLING proteins (small integrin-binding ligand, N-linked glycoproteins) (PMID 16776771). Interestingly, the descriptive terms, like *Glycoprotein*, *integrin-binding*, and *ligand* are all ranked high in the About Set and not present in the Full or Extra sets. Hence we might learn from the About Set that *osteopontin* is a SIBLING protein, but we can learn about other SIBLING proteins, like *DSPP* and *DMP-1* only from ExtraSent Set.

Despite a careful examination, we were not able to find any examples of key terms that were ranked significantly higher in Extra Set as compared to ExtraSent Set. More importantly, Extra Set gave several "false positives" (i.e. several highly ranked terms that were not associated with the gene) as compared to ExtraSent Set. This is in line with our original motivation for considering ExtraSent Set.

AboutSent Set. While ExtraSent was noticeably better than Extra Set, we found that this situation was not replicated when we compared About with AboutSent. In fact, when we compared the ranking of different types of key terms and across genes, the rankings of key terms given by About and AboutSent sets were very similar.

While there are some minor differences in the rankings by About Set and AboutSent Set, there was no noticeable pattern and our conclusion was that these provided very similar quality and type of information. In examining the differences between AboutSent and ExtraSent our observations suggested that there is a parallel to the situation we observed when comparing About with Extra.

FullSent Set. The documents in FullSent Set contain all sentences from the AboutSent and the ExtraSent sets. As we noted earlier, we felt that the About Set and AboutSent were not distinguishable, but the ExtraSent did provide better quality than Extra, as well as a useful but different kind of information from About. Preliminary analysis of the rankings of FullSent does indeed suggest that the advantages of these two sentence based documents were captured.

AboutTiFL Set. The reasons we considered the AboutTiFL Set are as follows: the title usually contains a short, yet concise, summary of the abstract, while the first sentence, as an introduction, together with the last sentence, as a conclusion, contain high level informative terms about the studies reported on the given gene. Thus, as we expected, we obtained most of the high level information related to the gene (such as *corepressor* for *Groucho*, *chromosomal translocation* for *Lmo2*, and *phosphoprotein* for *Opn*) but not highly relevant and detail oriented key terms. For example, *alkaline phosphatase activity* was ranked very low in the AboutTiFL for gene *Bmp2* while it ranked considerably high in the About Set. Similarly, other gene names, such as *osteocalcin* and *alp* which score highly in the About Set, do not appear in the top 150 key terms for the AboutTiFL Set. *WRPW* and *WD40* which are domains related to *Groucho* and extracted from the About Set are ranked low in AboutTiFL.

5.3 Conclusions

We have talked about differences among the Full, About, Extra, ExtraSent and AboutSent sets. We have seen how the Full Set does not distinguish extraneous information from important. By dividing the entire document set into About and Extra sets, we helped separate the two relevant types of information. More importantly, we have shown we can filter highly irrelevant information by considering the ExtraSent Set, which boosts ranks for potential interacting genes, similar or different genes, as well as domains and motifs relevant to the gene in question. We believe further investigation of FullSent versus About and ExtraSent sets is needed in order to determine if the About and ExtraSent sets give the most relevant key terms when used together, or if the FullSent set itself captures the information given by the two sets. On the other hand, if only high-level information is required, then we could restrict our query set to sentences in AboutTiFL.

One of key results of this work is that important concepts/key terms, to be associated with a given gene, can be extracted if we look in the right places for the particular type of concept. And hence, in our opinion, the Full Set (i.e. all abstracts retrieved by searching for a gene) is not the right place to extract key terms, whichever type of key term it is. In this context, we wish to point out that other systems appear to be using Full Set and not distinguish between different ways the gene is mentioned in an abstract.

Evaluating key terms is a challenging task, one of the many reasons being due to the lack of a gold set of terms relevant to specific genes. We are currently conducting an evaluation of key terms retrieved by eGIFT, based on ratings received from biologists, as well as by consulting manually created knowledge bases for genes to identify information which is captured/missed by eGIFT.

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.

Arek Gladki, Pawel Siedlecki, Szymon Kaczanowski, and Piotr Zielenkiewicz. 2008. e-LiSe–an online tool for finding needles in the '(Medline) haystack'. *Bioinformatics*, 24(8):1115–1117.

Jung-Jae Kim, Piotr Pezik, and Dietrich Rebholz-Schuhmann. 2008. MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline. *Bioinformatics*.

Neil R Smalheiser, Wei Zhou, and Vetle I Torvik. 2008. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. *Journal of Biomedical Discovery and Collaboration*, 3(1):2–11.

Catalina O Tudor, K Vijay-Shanker, and Carl J Schmidt. 2008. Mining the Biomedical Literature for Genic Information. In *Proceedings of the Workshop on Current Trends in BioNLP*, pages 28–29. Association for Computational Linguistics.