Elucidation of Protein Interaction via Google and Gene Ontology

B.V.Subba Rao[1], Dr.K.V.Sambasiva Rao[2]

[1]Associate Professor, Department of IT, P.V.P Siddhartha Institute of Technology, Vijayawada-7,India.
Email ID: bvsrau@gmail.com, Ph:9440109139.
[2]Principal, M.V.R. College of Engineering and Technology, Vijayawada, Krishna Dt., A.P, India
Email ID: principal@mvrcoe.ac.in, Ph:9440115556.

***Abstract*:** In the track of the growing quantity of biomedical text, there is a need for regular extraction of information to support biomedical researchers. Due to condensed biomedical information databases, the extraction cannot be done straightforward using dictionaries, so several approaches using associated rules and machine erudition have previously been proposed. Our work is motivated by the earlier approaches, but is novel in the sense that it combines Google and Gene Ontology for annotating protein connections. We got promising empirical results - 57.5% terms as valid GO annotations, and 16.9% protein names in the answers provided by our system ProG. The total error-rate was 25.6% consisting mainly of overly general answers and syntactic errors, but also including semantic errors, other biological entities and false information sources.

Keywords: Biomedical Literature, Data Mining, Gene Ontology, Google API.

1. Introduction

With the growing importance of precise and up-to-date databases about proteins and genes for research, there is a need for efficient ways of updating these databases by extracting information from biomedical research text [8], e.g. those indexed in MEDLINE. Examples of information resources containing such information are LocusLink, UniGene and Swiss-Prot for protein info and the Gene Ontology for semantic labels. Due to the huge and rapidly growing amounts of biomedical literature, the extraction process needs to be more automatic than previously. Current extraction approaches have provided promising results, but they are not sufficiently accurate and scalable. Methodologically all the suggested approaches belong to the information extraction field [3], and in the biomedical domain they range from simple automatic methods to more sophisticated, but slightly more manual, methods. Good examples are: Learning relationships between proteins/genes based on co-occurrences in MEDLINE abstracts [9] manually developed information extraction rules, information extraction (e.g. protein names) classifiers trained on manually annotated training corpora [12], and classifiers trained on automatically annotated training corpora.

A. Research Hypothesis

Internet Search Engines such as Google, Yahoo MSN, Bing, Alta Vista and Ask Me Search Engines are the world's largest readily available information sources, also in the biomedical domain. Based on promising results from recent work on using Google for semantic annotation of biomedical literature, we are encouraged to investigate if Google can be used to find protein interactions that match the Gene Ontology (GO). This leads to the hypothesis: Can Internet Search engines such as Google be used to annotate protein interactions in the Gene Ontology framework.
The rest of this paper is organized as follows. Section 2 describes the materials used, section 3 presents our method, section 4 presents empirical results, section 5 describes related work and section 6 describes conclusion and future work.

2. Materials

See fig. 1 for an overview of the system. As input for our experiments we used the following:
10 proteins that is already well-known to our biology experts. 37 verb-templates suggested by Martin.

A. Proteins

The following proteins were used as input to the system. Proteins user are 'EGF', 'TNF', 'CCK', 'gastrin', 'CCKBR', 'CREB' and 'CREM'.
In addition, each protein is also described by several other names or synonyms in the literature. E.g. gastrin is also known as 'g14', 'g17', 'g34', 'GAS', 'gast', 'gastrin precursor', 'gastrin 14', etc. So our biologists compiled a list of roughly 10 synonyms for each protein, giving us about 100 terms total to annotate.

B. Interaction Verbs

We selected our interaction verb templates from table 1 in. They had a list of 44 verbs, but we chose to use only 37 of these verbs. The reason for this is that we are focusing on simple statements like "gastrin activates", with the object of the verb following directly after the verb template. The following table shows the original list of verbs, with the removed ones in parenthesis.
Verb templates used are acetylates, activates, binds, blocks, bonds, degrades, hydrolyses, increases, interacts with, mediates, phosphorylates, reacts with, releases, stimulates, transforms, triggers, upregulates.

3. Our Approach

We have taken a modular approach where every sub module can easily be replaced by other similar modules in order to improve the general performance of the system. There are five modules in the system. The first one sets up the search queries, the second runs the queries against Google, the third one tokenizes the results, the fourth parses the tokenized text, and the fifth and last module extracts all the results and presents them to the human evaluators. See figure 1.

A. Data Selection
N (=100) protein names are combined with M (=37) verb templates, giving a total of N x M (3700) queries to run against Google.

B. Google
The queries are fed to the PyGoogle module which allows 1000 queries to be run against the Google search engine every day with a personal password key. In order to maximize the use of this quota, the results of every query are cached locally, so that each given query will be executed only once. If a search returns more than ten results, the resultset can be expanded by ten at a time, at the cost of one of the 1000 quota-queries every time. We decided to use up to 30 results for each query in this experiment.

C. Tokenization
The text is tokenized to split it into meaningful tokens, or "words". We use a simple WhiteSpaceTokenizer from NLTK, where every special character (like ( ) " ' - , and .) is treated as a separate token.

D. Parsing
Each returned hit from Google contains a "snippet" with the given query phrase and approximately ten words on each side of it. We use some simple regular grammars to match the phrase and the words following it. If the next word is a noun it is returned. Otherwise, adjectives are skipped until a noun is encountered, or a "miss" is returned.

E. Expert Evaluation
The results were merged so that all synonyms were treated as if the main protein name had been used in the original query. Then the results were put into groups (one group for each protein-verb pair) and sorted alphabetically within that group. These results were then presented to the biologists, who evaluated the usefulness of our results from Google.

4. Empirical Results

Fig. 2 and 3 show the results. The first one shows that more than half of the extracted terms were terms that could be used to annotate the given protein around one fifth of the results contained an identifiable protein name that could be stored as a protein-protein interaction.

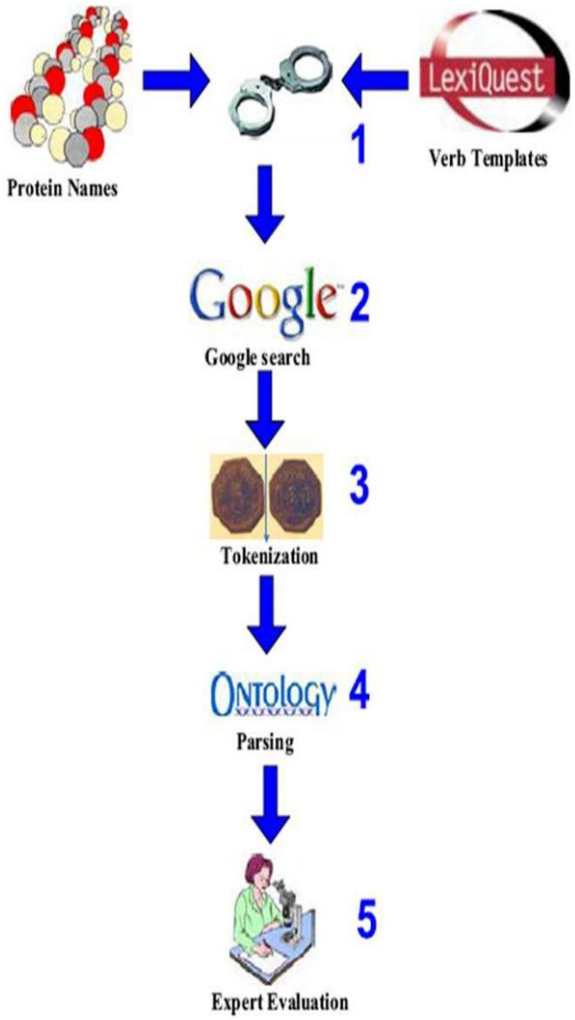

Fig. 1. Overview of our Approach (named ProG) according to the Gene Ontology (GO).

Only one quarter of the terms were deemed not useful. The different kinds of "not useful"-errors can be read out of fig. 3.

5. Related Work

Our specific approach was on using Google and Gene Ontology for annotating protein interactions. We haven't been able to find other work that does this, but the closest are Dingare et al., that uses results from Google search as a feature for a maximum entropy classifier used to detect protein and gene names [5, 6] and our previous work on semantic annotation of proteins (i.e. tagging of individual proteins, not their GO relation). Google has also been used for semantic tagging outside of the biomedical field, e.g. in Cimiano and Staab's PANKOW system [2] and in [4, 7, 10, 11].
A comprehensive overview of past methods for protein-related information extraction is provided in.

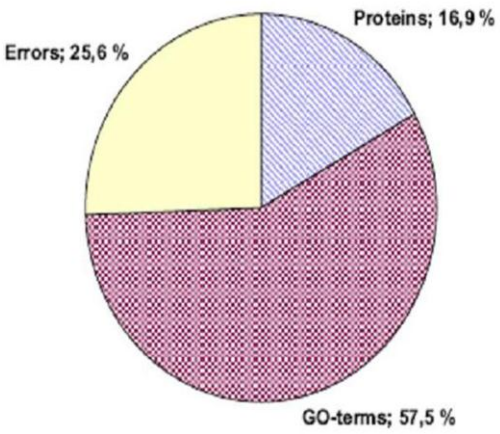

Fig. 2. Main Results

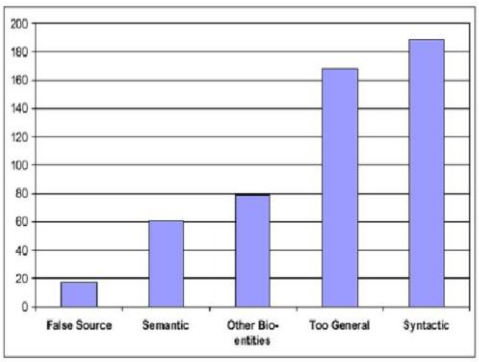

Fig. 3. Breakdown of Errors

6. Conclusion And Future Work

This paper presents a novel approach - ProG - using Google to find semantic (GO-) annotations for specific proteins. We got empirically promising results - 57.5% semantic annotation classes, and 16.9% protein names in the answers provided by ProG. This means that 74.4% of the results are useful. This encourages further work, possibly in combination with other approaches (e.g. rule based information extraction methods), in order to improve the overall accuracy. In the similar task of protein name identification, recently presented precision scores ranges from 70 to 75% [1]. Hopefully, more advanced methods will greatly reduce the number of errors (useless information), which is currently at 25.6%. Disambiguation is another issue that needs to be further investigated, because sometimes different search-results are really just one single identity, because of synonyms and acronyms for example. Other opportunities for future work include:

– Improve tokenization. Just splitting on whitespace and punctuation characters is *not* good enough. In biomedical texts non-alphabetic characters such as brackets and dashes need to be handled better.

– Search for other verb templates using Google. E.g. Which templates give the best results, and what about negations ("does not activate ...")

– Investigate whether the Google ranking is correlated with the accuracy of the proposed semantic tag. Are highly ranked pages better sources than lower ranked ones?

– Test our approach on larger datasets, e.g. using *all* the returned results from Google.

– Combine this approach with more advanced natural language parsing techniques in order to improve the accuracy.

– In order to find multiword tokens, one could extend the search query *("X activates")* to also include neighboring words of X, and then see how this affects the number of hits returned by Google. If there is no reduction in the number of hits, this means that the words are "always" printed together and are likely constituents in a multiword token. If you have only one actual hit to begin with, the certainty of the previous statement is of course very weak, but with increasing number of hits, the confidence is also growing. – In this experiment very crude Part Of Speech (POS) tagging is done, so our results can be seen as a baseline for this kind of experiment. In the future we want to improve the results, for example by utilizing better grammars, and more advanced natural language understanding techniques.

References

[1] Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and YukWah Wong. Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Journal Artificial Intelligence in Medicine: Special Issue on Summarization and Information Extraction from Medical Documents (Forthcoming)*, 2004.

[2] Philipp Cimiano and Steffen Staab. Learning by Googling. *SIGKDD Explorations Newsletter*, 6(2):24–34, December 2004.

[3] J. Cowie and W. Lehnert. Information Extraction. *Communications of the ACM*, 39(1):80–91, January 1996.

[4] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. SemTag and seeker: bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the Twelfth International World Wide Web Conference, WWW2003*, pages 178–186. ACM, 2003.

[5] Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, and Beatrice Alex. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. In *Proceedings of the BioCreative Workshop*, March 2004.

[6] Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. Submitted to BMC Bioinformatics, 2004.

[7] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised Named- Entity Extraction from the Web: An

Experimental Study. Submitted to Artificial Intelligence, 2004.

[8] Jun ichi Tsuji and Limsoon Wong. Natural Language Processing and Information Extraction in Biology. In *Proceedings of the Pacific Symposium on Biocomputing 2001*, pages 372–373, 2001.

[9] Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski, and Eivind Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28, May 2001.

[10] Vinay Kakade and Madhura Sharangpani. Improving the Precision of Web Search for Medical Domain using Automatic Query Expansion. Online, 2004.

[11] Udo Kruschwitz. Automatically Acquired Domain Knowledge for ad hoc Search: Evaluation Results. In *Proceedings of the 2003 Intl. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE'03)*. IEEE, 2003.

[12] B.V.Subba Rao, Dr.K.V.Sambasiva Rao, Semantic Explanation of Biomedical Literature using Google. *In proceedings International conference on Websciences-2009, pages 452-457.*

Author's Biographies

B.V.Subba Rao, presently working as Associate Professor in P.V.P Siddhartha Institute of Technology Vijayawada, affiliated to Jawaharlal Nehru Technological University. He received his M.Tech degree with distinction in Computer Science and Engineering from Acharya Nagarjuna,University. He received Gold medal in his post graduate studies. He is pursuing Ph.D in Computer Science and Engineering at Acharya Nagarjuna University, Guntur. He has guided 30 post Graduated and 40 graduate projects. He has published 4 papers (National / Conference Proceedings) and has Academic participation in 24 International / National Seminars / workshops and Conferences. He is a member of Computer Society of India (CSI), Association for Computing Machinery (ACM), and Indian Society for Technical Education (ISTE). His current research interests are in the areas of Artificial Intelligence, Natural Language Processing, Information Retrieval systems and Bioinformatics.

Dr.K.V.Sambasiva Rao, presently working as a Principal of MVR College of Engineering and Technology, Paritala. He pursued his M.E from BITS, Pilani and Doctorate from IIT Delhi. He has a total of 21 years of rich experience comprising teaching, research and industry. He has published 4 books, 18 papers in international and national journals. He has conducted numerous national conferences, workshops with the support of AICTE, DST and other government bodies. He has given more than 50 seminar talks at various technical institutions. He has guided 25 Masters level projects and is research director for 11 Ph.D candidates. His biography was included in MARQUI'S INTERNATIONAL, New Jersey, USA "Who is who in the World" in the year 1999 and was awarded "Outstanding achievement award" by International Biography Centre, Cambridge, UK. He is the life member of 3 professional bodies.