Finding Gene Relations in Scientific Abstracts

André Montevecchi, Ismael Santana, Douglas Gonçalves, Wander Júnior, Glívia Barbosa, Gustavo Ornelas, Ulisses Fernandes, Saulo Pinto

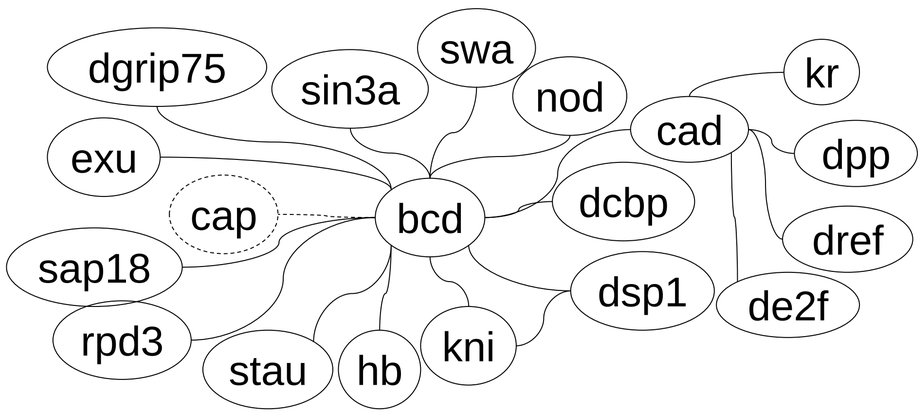
1 Introduction

The amount of scientific literature published daily is growing fast. There are biomedical databases storing full papers, abstracts, and references in order to facilitate information dissemination and retrieval. By far, *PubMed* is the most accessed of such databases. By January, 2007, more than 90,000 searches were performed to retrieve information from more than 14 million citations available in the PubMed database [1]. In such a wealthy of information available is very difficult for a researcher to find the information she/he needs in a way to facilitate visualization and, mainly, the search for relationships among entities being investigated. The establishment of relations by means of research abstracts mining is a recent but not a new idea [2, 3, 4]. However, the whole methodology partially described here is a novel one. This text presents initial results of the basic step of an ongoing effort to implement a "plausible causal mechanism finder" methodology to reconstruct gene causal networks using information publicly available in the form of paper abstracts.

2 The Methodology

The search for gene relations is started using a gene identifier, usually a common gene symbol, hereinafter named the "target" (e.g. "bcd", for the bicoid gene), and the organism's name (e.g. *Drosophila melanogaster*). Every target aliases, including its full name (description), are retrieved from a local aliases repository previously downloaded from NCBI's Gene database [5] and a query string is built in the following format: (id_1 OR id_2 OR… OR id_n) AND (OrganismName1 OR OrganismName2), where id_i is a target's alias. For the preceding example, the following string is built: (bcd OR bic OR bicoid) AND (Drosophila OR melanogaster). After that, PubMed is remotely accessed by using the Entrez e-Utilities [6] and the query is sent to it. The abstracts satisfying the query are retrieved together to their ids and titles. Now, for each title/abstract we scan it for known gene symbols. Each unseen gene symbol that is found in a title/abstract generates a relationship between it and the target. The relationship itself, the title, and the abstract (and its PubMed id) that generate the relationship are stored locally for further processing. The process continues over all abstracts retrieved from PubMed and it's recursively repeated with each gene related to the target being a new target until no unseen abstract is returned by Entrez.

3 Results



The methodology briefly described here is suited to build a network of genes by means of their co-occurrence in an abstract. We run our .NET/Java software to the bcd gene (*D. melanogaster*). The inferred bcd net has 88 nodes (genes), including itself. The relations were detected from 278 abstracts, totalizing 607 associations between bcd and another gene. A small fraction is shown in the figure below. We verified the retrieved associations centered on bcd by reading the abstracts and found out that 71% of the relations involving it are correctly detected in at least one abstract. Relations could be anyone: from regulatory to morphogenetic interactions.

4 Conclusion

The initial results presented here are promising since we have considered only co-occurrence to infer that a relation is present in an abstract. The introduction of Machine Learning techniques will improve the overall performance. Currently we are implementing a Web interface to allow the network visualization. The user will be capable to visualize the relations and to get access to the most important abstracts and the full papers if they are available through the Internet..

5 References

[1]http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html.
[2] Stephens, M. Palakal, M., Mukhopadhyay, S., Raje, R. "Detecting Gene Relations from Medline Abstracts". Pacific Symp. on Biocomputing, 2001.
[3] Jenssen, T. K., Laegreid, A., Komorowski, J., Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, May;28(1):21-8, 2001
[4] Šarić, J., Jensen, L. J., Ouzounova, R., Rojas, I., Bork, P. Extraction of regulatory gene/protein networks from Medline. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2005.
[5] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene.
[6]http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.
[7] http://www.pagex.com/webtools/stopwords.cfm and http://adsabs.harvard.edu/abs_doc/stopwords.html.