Biomedical Literature Mining

Mario Fruzangohar

In the fulfilments of the degree of

Doctor of Philosophy

A thesis by prior publications submitted to

Discipline of Genetics

School of Biomedical and Health Sciences

The University of Adelaide

February 2014

Table of Contents

Acknowledgments

I first wish to thank my principal supervisor Prof. David Adelson who is one of the most patient people I have ever met, always welcoming me, even when I had ideas of weird experiments! Thank you David for the enduring support you have given me throughout my candidature. I would also like to thank my co-supervisor Prof. Hong Shen from computer science school. I also truly acknowledge the help and support I have received from Dr. Esmaeil Ebrahimi and also his precious experiences he shared with me.

This research project would not have been possible without the bacterial data providedbymycolleagues at the Research Centre for Infectious Diseases, namely Dr. David Ogunniyi, Dr. Layla Mahdi and Prof. James Paton. I am grateful to all of them for their time and patience.

I must not and cannot forget the significance of the friendships I have made during my candidature here in University of Adelaide. I do not want to miss anyone by naming people individually. I have never overlooked the value of a friendly chat, motivating me through the rest of the day.

Finally and most sincerely, I wish to deeply thank my precious family and friends who gave me the strength and courage to continue my studies by their support and love.

Abstract

Thousands of biomedical articles are published every year containing many newly discovered biological interactions and functions. Manually reading and classifying this information is a difficult and laborious task. Literature mining contains mechanisms and tools to automate the process of extracting biological relationships, storing them in biological databases and finally analyse and present them in a biological meaningful way. In the first stage of literature mining, articles are parsed and get segmented, sentences separated, tokenized and finally annotated by part of speech tags (POS).

POS tagging is the most challenging part because the training corpus is relatively small compared to the large number of biological names therefore limiting the lexicon. There are a number of solutions to address this problem including extending the lexicon manually or using character features of the word. There is no empirical comparison between different solutions. So we developed a complete list of tools including article parser, segmentation, sentence detector, sentence tokeniser, POS tagger and finally noun phrase detector using JAVA and PostgreSQL technologies. We tailored these tools for biomedical texts, and empirically compared them with other tools and we demonstrated increased efficiency of our tools compared to others.

Once biological relationships are extracted they are ready to be stored in databases to be used and shared by others. There a wide range of databases that store annotation data related to genes, proteins and other biological entities. Among them Gene Ontology annotation database is the key database that connects all the other biological entities through a standard vocabulary together. In fact a Gene Ontology (GO) is a controlled vocabulary to annotate proteins based on their molecular function, biological process and cellular components. There are a number of public databases that provide data regarding GO and GO-protein relationships. We collected all relevant data from several public databases and built our specialized updatable GO database on the PostgreSQL platform.

GO classification in a particular sample of genes (up/down regulated) or whole genome of a species can reveal the biological mechanisms related to its activity. Moreover, comparing the GO classification of a species under different biological conditions can elucidate its biological pathways, which can result in the discovery of novel genes to be used in therapies.

We developed a web server using the PHP MVC framework connected to our specialized GO database. In this web server we developed novel visual and statistical methods to perform GO comparisons among multiple samples and genomes.

We also included transcriptome based gene expression levels in GO analysis, resulting in novel meaningful biological reports. This also made comparison of whole genome gene expression across multiple biological conditions possible.

Furthermore, we devised a method to dynamically construct and visualize GO regulatory networks for any gene set sample. Such a network can reveal regulatory relationships between genes helping to explain the correlated expression of genes. The topology of such a network classifies genes based on their connections, and can be used as a new method to detect important genes based on their function as well as their connectivity in the network.

We demonstrated the efficiency of our developed methods in our web server by several case studies using previously published transcriptome data.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Mario Fruzangohar *Date*

List of Publications

1. Fruzangohar M, Kroeger TA, Adelson DL (2013) Improved part-of-speech prediction in suffix analysis. PloS one 8: e76042.

2. Fruzangohar M, Ebrahimie E, Ogunniyi AD, Mahdi LK, Paton JC, et al. (2013) Comparative GO: A Web Application for Comparative Gene Ontology and Gene Ontology-Based Gene Selection in Bacteria. PloS one 8: e58759.

3.Fruzangohar M, Ebrahimie E, Adelson DL (2014)Application of Global Transcriptome data in Gene Ontology Classification and Gene Ontology Interaction Networknetwork. Manuscript Prepared