MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)
© 2022 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220284

Exploring Molecular Mechanisms Within Biomedical Literature

Nicole Glendinning[a], Christopher Hawthorne[a], Stephanie Beck[a], Guillermo Lopez-Campos[a]

[a] Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom

Abstract

Text mining of the biomedical literature enables vast quantities of information to be extracted and summarised. Here we describe an updated and improved version of previous methodology for the analysis of gene and protein biomarkers that enables the use of the newer Pubtator Central annotations, based in full text, improving the performance using a local SQLite database, that reduces the running time and resources required to perform the analyses facilitating its use in any computer, and expands its capabilities to enable the retrieval and analysis of chemical and metabolic biomarkers.

Keywords:

Computational Biology, Data Mining, Biomarkers

Introduction

Biobliographic resources represent an important source of biomedical data and publications represent the main avenue to share advances in science. Biomedical text mining is an incredibly valuable research technique because of its use in deciphering important information and concepts from large amounts of scientific literature. Research in this area has produced a variety of algorithms and solutions that are able to identify and extract biomedical terms from scientific texts. An example of these tools are PubTator and its successor PubTator Central [1] which are freely available and highly regarded in supporting biomedical text mining through providing information on various biomedical concepts associated with PubMed articles and abstracts. We have previously developed an approach for the analysis of biomarkers based on the use of PubTator flat annotation files to explore similar and different molecular mechanisms associated with the retrieveal and analysis of genes derived from different PubMed queries [2]. However, this approach relied in loading the large annotations files into R for the downstream analyses and was limited to the analysis of genes and proteins not considering chemicals and metabolites as part of the analyses.

In this work, we are presenting a modified and updated version of the methodology implementing new data structures and updating the annotation sources (using Pubtator Central full text annotations that replace old PubTator annotations and were discontinued in February 2020) and expands the analysis to chemicals and meabolites applying them in the analysis of biomarkers for preeclampsia, a serious condition potentially fatal if left untreated that affects 5 to 10% pregnancies globally.

Methods

We initially updated our R script to be able to read and manage the newer PubTator Central annotation files (the structure is different to previous PubTator files), including the newer "Cell line" annotations for the comparative analyses across different Pubmed queries. To improve the performance of our approach, we replaced the reading the annotation files into R by a SQLite database containing one table per annotation file (Cell line, Chemical, Disease, Gene, Mutation, and Species). To generate this database we initially adapted the R package "pubtatordb" [3] to ensure its compatibility with the newer format and contents, which includes cell lines, of PubTator Central annotations. Finally, we assessed the performance of the new solution using the "profvis" package in R. For this purpose we defined a Pubmed query to retrieve and analyse chemicals (chemical2pubator) using alternatively the new database (~16.2GB) approach, reading the "chemical" flat file (~5.2GB) into R and reading the full annotation (Bioconcept2pubtator) flat file (~16.6GB) into R on an iMac Pro with 128GB RAM.

Genes and metabolites were both analysed through the text mining approach of the script, with the extraction of metabolites and genes from literature from PubMed alongside pathway enrichment on the Reactome platform. A combined analysis approach (a combination of both genes and metabolites) was employed for the potential influence of these two targets on preeclampsia. The statistical test, Fisher's Exact test was carried out with a Bonferroni adjusted statistical threshold of 0.05.

Results

The presented version of the tool is now able to use the latest PubTator Central annotation files, while maintaining the compatibility with the older PubTator annotation files (ceased in February 2020) and handles the annotation contents either in the format of a local SQLite database (preferred) or the original flat files.

The results from the performance assessment (table 1.) showed a great improvement derived from the use of the SQLite database. As expected the SQLite database approach greatly improved the performance both in execution time and memory usage.

	Chemical2Pubtator (old version script)	Bioconcept2Pubator (old version script)	SQLite (new version script)
RAM Memory (MB)	21188.1	28859.9	963.7
Time (ms)	177970	544660	39580

Table 1. *Performance assessment using using "profvis" function in R for the same query*

From the script, 893 metabolite and chemicals and 1271 genes were extracted from the literature and assumed to have an association with pre-eclampsia. Pathway enrichment was carried out for these biomarkers. The pathway enrichment from Reactome produced, 293 and 30 were found to be statistically significant for both genes and metabolites. For the combined, roughly 50% of the pathways had a statistically significant association.

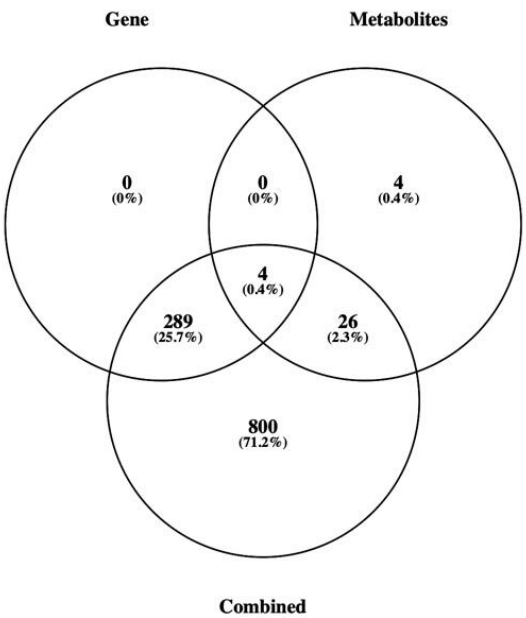

Figure 1–Venn diagram displaying the overlap from the analyses using pathway enrichment analysis for genes, metabolites and the combination of both.

Conclusions

The incorporation of metabolite analysis has further modified and expanded the potential of this script. From the retrieval of biomarkers, pathway enrichment can occur which may give further insight into understanding the mechanisms underlying the condition. This project has expanded and enhanced the capabilities of the previous version, providing the possibility of retrieving and analysing chemicals and metabolites. The use of individual analysis of genes and metabolites shows little overlap between them in terms of pathway while the combination of both provides a much deeper insight in the potential mechanisms.

The modifications carried out have improved the performance reducing the time and resources required to run the analyses, allowing for them to be easily completed in any computer equipped with R without requiring large amounts of RAM. This is especially true for complex analyses involving different elements (genes, chemicals, mutations, ...) that would require either the use of the complete annotation file or repeated analyses using the different individual annotation files.

This new version of the script was used to analyse the literature looking for biomarkers for pre-eclampsia identifying metabolites that complement and enrich previous analyses based on genes and proteins and providing mechanistic insight about the molecular processes involved in this serious condition.

References

[1] CH Wei, A Allot, R Leaman, Z Lu PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res*. **47(W1) (**2019), W587-W593

[2] G Lopez-Campos, E Bonner, L McClements, An Integrative Biomedical Informatics Approach to Elucidate the Similarities Between Pre-Eclampsia and Hypertension. *Stud Health Technol Inform.* **264** (2019), 988-992

[3] https://cran.r-project.org/web/packages/pubtatordb/index.html

Address for correspondence

Guillermo Lopez Campos, Address: Wellcome-Wolfson Institute for Experimental Medicine, 97 Lisburn Road, Belfast, BT9 7BL, United Kingdom. Email: g.lopezcampos@qub.ac.uk..