Exploring the Feasibility of Automating Biocuration for Neuropharmacology and Zebrafish (*Danio rerio*)

Directed Independent Study: Spring 2022

University of North Florida, School of Computing

Victoria Leventman, Dr. Indika Kahanda

April 27, 2022

Contents

Abstract

The growing number of published biomedical articles stored in literature databases is far outpacing the rate at which bio-researchers can manually examine and annotate the literature to best meet their research needs. The curation of articles relevant to various biological subjects accommodates extending further research and fostering new developments and hypotheses as biologists enhance their domain knowledge. In this study, we investigated the feasibility of biocuration for the neuropharmacology field, with a focus on the model organism zebrafish (*Danio rerio*), and its potential to be utilized for automated annotations. The study first verified that literature related to genes in the GABA pathway, the drug Ivermectin, and zebrafish were readily available in biomedical literature databases such as PubMed and PubMed Central. This verification was achieved by programmatically developing an Information Retrieval pipeline and employing three E-Utilities provided by the National Center for Biotechnology Information (NCBI): ESearch, ESummary, and ELink.

Once a filtered subset of top-10 biomedical abstracts from the query results for "GABA AND Ivermectin NOT covid-19" were manually annotated, it was found that four abstracts had higher observed relevancy due to the presence of particular biological keywords and phrases. These findings are significant as they establish how bio-ontologies specific to neuropharmacology can be used with a dictionary mapping tool such as ConceptMapper to automatically generate valuable annotations. With information retrieval and annotated literature, it can evolve to become a knowledge base on neuroactive drug discovery with zebrafish that is made available to bio-researchers.

1 Introduction

Scientific articles are getting published in literature databases at an exponential rate (8). With this fast growth of information available online, it is difficult for biocurators and bio-researchers to manually sort through, group, and annotate millions of articles. For various fields of biology, biocuration aims to curate relevant information from literature related to biological entities (e.g., zebrafish - *Danio rerio*, which is a freshwater fish) and populate knowledge bases such as UniProt[1]. Knowledge bases are helpful and applicable to bio-research as they contain the most meaningful information for subjects of interest. The motivation behind our study was programmatically establishing an Information Retrieval (IR) pipeline for future bio-researchers

[1] https://www.uniprot.org/help/uniprotkb

in the growing area of neuropharmacology. In more detail, the field of neuropharmacology refers to the study of drugs and their effects on the nervous system within organisms. As our IR pipeline retrieves relevant biomedical literature, it would also address the efficiency of the curation process through the aggregation of this literature and thus allowing researchers to improve their understanding of biological topics.

2 Related Work

The topics of information retrieval and biocuration were discussed in the KinDER paper by Dopp et al. (1). In this paper, the authors focused on the process of auto-annotating and curating biomedical journal articles (from literature databases) about human protein kinases, which is a biological entity that has use cases in drug design. During this process, they aimed to form a knowledge base of relevant information on kinases. The authors created the KinDER (Kinase Document Extractor and Ranker) pipeline, which consisted of two components: retrieving the literature and classifying the relevancy of documents via machine learning models. Additionally, the topic of literature analysis for the neuropharmacology field was discussed in the paper by Yeung et al. (7). The authors analyzed word frequencies, along with the co-occurrences of words, and the citation counts of 40,000 and more neuropharmacology articles from the Web of Science Core Collection[2] database. Their analysis aimed to determine which research topics in the field were popular among researchers currently. Based on their results from the articles' titles and abstracts, which were visualized in a bubble map, the authors found that "pathway", "activation", "inhibition", "neuron", and "receptor" were in the list of top 10 words that appeared most often.

The overall strengths of the existing literature are how they indicate that it is feasible to simulate the authors' methods of programmatically extracting literature, obtaining word frequencies, and aggregating relevant articles on biological topics such as neurotransmitter pathways and neuroactive drugs within the area of neuropharmacology.

3 Data and Tools

In our study, the biological topics of interest were the model organism Zebrafish, the GABA pathway (which is a neurotransmitter pathway), genes

[2] https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/

in that pathway, and the drug Ivermectin. There were two primary data sources for the retrieval of the biomedical literature, with the first one being the PubMed[3] database, and the second one being the PubMed Central[4] (PMC) database. The PubMed database solely contains abstracts, while the PMC database contains full-text articles. For the tools to carry out this retrieval and extraction of knowledge from the databases, we used Biopython (5) to programmatically obtain the literature with three Entrez Programming Utilities (E-utilities): ESearch, ESummary, and ELink. E-utilities offer a wide variety of functionalities for accessing the NCBI Entrez databases. Two of the Entrez databases are PubMed and PMC, for example.

4 Methodology and Experimental Setup

To describe the data pipeline in more detail, we have this Information Retrieval pipeline (Fig. 1) where inputs go in, some processing is completed internally, and then the outputs are received on the other end. The inputs within the pipeline were search queries for a database, whether that was PubMed or PMC. The outputs of our pipeline were the top 10 articles per query. The data processing itself within the pipeline corresponds to the previously mentioned E-utilities: ESearch, ESummary, and ELink. The E-utilities offer a plethora of information, but within our pipeline, there were specific use cases. ESearch allowed us to retrieve article IDs for the provided input queries, and the IDs were sorted by relevancy based on the respective Best Match algorithms used in PubMed and PMC (2; 6). ESummary allowed us to extract the publication date (PubDate) per article. Then, with ELink, there were connections, or links, between articles that helped to determine the citation counts per article. Here, citation counts refer to the number of times an article has been cited by other publications.

In the pipeline, a filtering step was done, and this step is crucial to getting the end results to a more manageable size. The filtering criterion was at least 25 citation counts for each article, and this criterion of 25 citation counts or more is significant because it signifies that the article is reputable. Reputability is important when bio-researchers examine these publications for their own work, in addition to future research and experiments.

For our experimental setup, five genes within the GABA pathway were considered as the starting place: GABRA1, GABRA2, GABRA3, GABRA4, and GABRA5, and they refer to our gene names. The motivation behind

[3] https://pubmed.ncbi.nlm.nih.gov/

[4] https://www.ncbi.nlm.nih.gov/pmc/

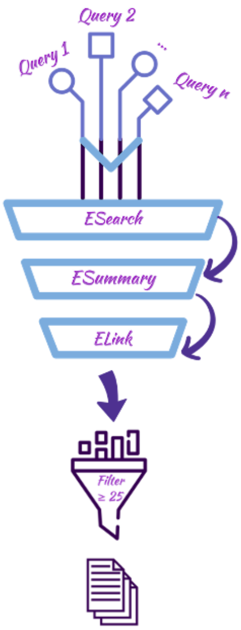
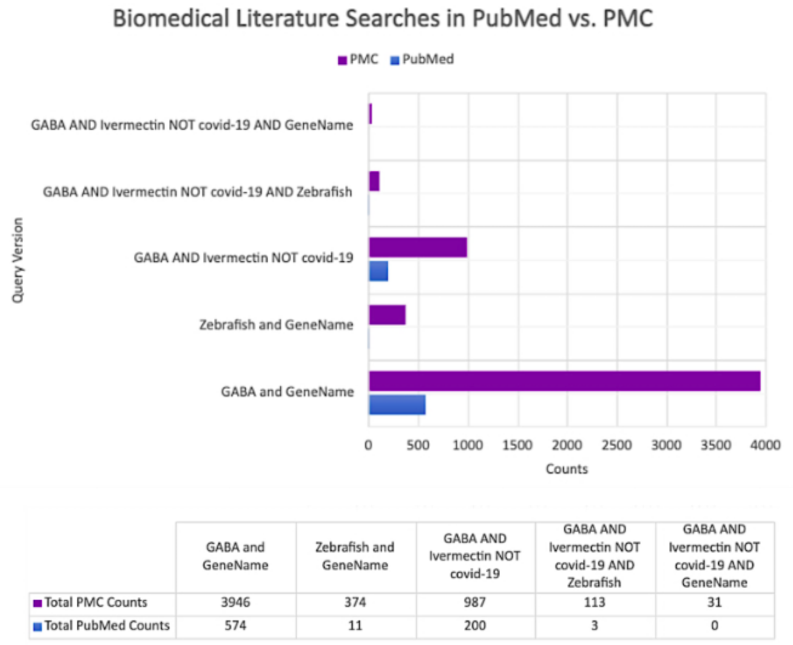


Fig. 1. Information Retrieval pipeline with filtering

the gene names is to drill down and see if there are articles discussing Ivermectin's effect on one or more genes in the GABA pathway. Furthermore, six queries in total were the inputs in the IR pipeline, and they are listed as follows and ordered from least specific to most specific:

1. "GABA and GeneName"
2. "Zebrafish and GeneName"
3. "GABA AND Ivermectin NOT covid-19"
4. "GABA AND Ivermectin NOT covid-19 AND Zebrafish"
5. "GABA AND Ivermectin NOT covid-19 AND GeneName"
6. "GABA AND Ivermectin NOT covid-19 AND Zebrafish AND GeneName"

The reasoning for excluding the term "COVID-19" is that it was a confounding factor and not the main focus of the biological research within this study.



	GABA and GeneName	Zebrafish and GeneName	GABA AND Ivermectin NOT covid-19	GABA AND Ivermectin NOT covid-19 AND Zebrafish	GABA AND Ivermectin NOT covid-19 AND GeneName
■ Total PMC Counts	3946	374	987	113	31
■ Total PubMed Counts	574	11	200	3	0

Fig. 2. Literature search results gathered from the pipeline

5 Results and Conclusions

From the results we gathered (see Fig. 2), the overall pattern we saw was that the PubMed Central (PMC) database returned more results for the biomedical literature searches compared to PubMed. Particularly, for queries 1-5, PMC returned noticeably more results (in total) compared to PubMed. Due to its specificness by having four main terms (GABA, Ivermectin, Zebrafish, GeneName), the 6th query returned no results in both biomedical literature databases.

In the conclusion of our study, Julia Gabel, an undergraduate researcher from the Department of Biology, manually annotated 10 PubMed abstracts from the "GABA and Ivermectin NOT covid-19" query. From the annotations, we identified how there were key bio-terminologies in certain abstracts, which led them to have higher observed relevancy compared to the remaining abstracts. The significance of our findings is that it is feasible to automatically generate annotated documents by using a dictionary mapping tool such as ConceptMapper and standard vocabularies of biology terms (i.e., bio-ontologies). Then, we can develop a gold-standard dataset of annotated articles labeled as relevant or irrelevant to the subject matter, such as

Ivermectin's effect on neurotransmitter pathways. Machine learning models trained on the gold standard can then be used to classify the relevancy of new articles based on the presence of specific keywords and phrases. As the final step, the relevant documents would get curated to create this valuable knowledge base that bio-researchers can utilize.

6 Future Work

Based on the specific hypotheses that the bio-researchers have, it will be beneficial to obtain from the researchers a list of keywords that represent what their ideal article would contain. The reasoning behind doing this on the programming end is to better understand the information needs and connect them to the queries that would be tested out within the pipeline. Furthermore, for the IR pipeline, additional data processing can be performed on the articles to determine the relative positioning of the terms ”GABA” and ”Ivermectin” per sentence or paragraph. If the terms are located close to each other numerous times throughout an article, it would indicate that this article contains relevant information on the direct connections between the two biological entities rather than addressing them separately.

7 Acknowledgements

We would like to thank Dr. Marie Mooney and Julia Gabel from the Department of Biology for their direction, guidance, and support.

8 References

1. Dopp, Daniel, et al. ”KinDER: A Biocuration Tool for Extracting Kinase Knowledge from Biomedical Literature.” *Proceedings of the BioCreative VI Workshop*, Oct. 2017, pp. 51-55.
2. Fiorini, Nicolas, et al. ”Best Match: New relevance search for PubMed.” *PLoS biology*, vol. 16, no. 8, Aug. 2018, pp. e2005343.
3. Howe, Doug, et al. ”Big Data The Future of Biocuration.” *Nature (London)*, vol. 455, no. 7209, Sept. 2008, pp. 47–50.
4. International Society for Biocuration. ”Biocuration: Distilling Data into Knowledge.” *PLoS Biology*, vol. 16, no. 4, Apr. 2018, pp. e2002846.
5. Peter J. A. Cock, et al. ”Biopython: freely available Python tools for computational molecular biology and bioinformatics.” *Bioinformatics*, vol. 25, no. 11, Jun. 2009, pp. 1422–1423.
6. ”PMC User Guide.” *National Center for Biotechnology Information*, U.S. National Library of Medicine.
7. Yeung, Andy Wai Kan, et al. “When Neuroscience Meets Pharmacology: A Neuropharmacology Literature Analysis.” *Frontiers in Neuroscience*, vol. 12, Nov. 2018, pp. 1-7.
8. Zhu, Peiyan, et al. ”Mining meaningful topics from massive biomedical literature.” *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2014, pp. 438-443.