Title

BioTextQuest v2.0: An evolved tool for biomedical literature mining and concept discovery.

Authors

Theodosiou, Theodosios
Vrettos, Konstantinos
Baltsavia, Ismini
et al.

Publication Date

2024-12-01

DOI

10.1016/j.csbj.2024.08.016

Peer reviewed

Computational and Structural Biotechnology Journal 23 (2024) 3247–3253

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

BioTextQuest v2.0: An evolved tool for biomedical literature mining and concept discovery☆

Theodosios Theodosiou [a], Konstantinos Vrettos [a], Ismini Baltsavia [a], Fotis Baltoumas [b], Nikolas Papanikolaou [a], Andreas N. Antonakis [a], Dimitrios Mossialos [c], Christos A. Ouzounis [d], Vasilis J. Promponas [e], Makrina Karaglani [f], Ekaterini Chatzaki [f], Sven Brandau [g], Georgios A. Pavlopoulos [b], Evangelos Andreakos [h], Ioannis Iliopoulos [a,*]

[a] Division of Basic Sciences, University of Crete Medical School, Heraklion 71110, Greece
[b] Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari, Athens 16672, Greece
[c] Department of Biochemistry and Biotechnology, University of Thessaly, 41500 Larissa, Greece
[d] Biological Computation & Computational Biology Group, AIIA Lab, School of Informatics, Aristotle University of Thessalonica, 57001 Thessalonica, Greece
[e] Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, Nicosia 1678, Cyprus
[f] Medical School, Democritus University of Thrace, 68100 Alexandroupolis, Greece
[g] Experimental and Translational Research, Department of Otorhinolaryngology, University Hospital Essen, Essen, Germany
[h] Center for Immunology and Transplantation, Biomedical Research Foundation Academy of Athens, Athens, Greece

ARTICLE INFO

Keywords:
Biomedical literature mining
Concept discovery

ABSTRACT

The process of navigating through the landscape of biomedical literature and performing searches or combining them with bioinformatics analyses can be daunting, considering the exponential growth of scientific corpora and the plethora of tools designed to mine PubMed(®) and related repositories. Herein, we present BioTextQuest v2.0, a tool for biomedical literature mining. BioTextQuest v2.0 is an open-source online web portal for document clustering based on sets of selected biomedical terms, offering efficient management of information derived from PubMed abstracts. Employing established machine learning algorithms, the tool facilitates document clustering while allowing users to customize the analysis by selecting terms of interest. BioTextQuest v2.0 streamlines the process of uncovering valuable insights from biomedical research articles, serving as an agent that connects the identification of key terms like genes/proteins, diseases, chemicals, Gene Ontology (GO) terms, functions, and others through named entity recognition, and their application in biological research. Instead of manually sifting through articles, researchers can enter their PubMed-like query and receive extracted information in two user-friendly formats, tables and word clouds, simplifying the comprehension of key findings. The latest update of BioTextQuest leverages the EXTRACT named entity recognition tagger, enhancing its ability to pinpoint various biological entities within text. BioTextQuest v2.0 acts as a research assistant, significantly reducing the time and effort required for researchers to identify and present relevant information from the biomedical literature.

1. Introduction

The exponential growth of biomedical literature and the proliferation of repositories housing biological data present formidable challenges for researchers. PubMed® alone contains over 35 million MEDLINE entries, underscoring the sheer volume of information contained in publications from peer-reviewed journals. Furthermore, starting from January 2023 and onwards, the service also includes selected preprints from sources such as bioRxiv, medRxiv, and Research Square, further increasing its content (link: https://www.ncbi.nlm.nih.gov/pmc/about/nihpreprints/). At the same time, data from publications is frequently coupled with annotations derived from genomic, proteomic, and clinical repositories, presenting new opportunities and challenges. Overall, the volume of biomedical information demands

☆ BioTextQuest v2.0 is available at: http://bioinformatics.med.uoc.gr/shinyapps/app/biotextquest

* Corresponding author.
E-mail address: ioannis@uoc.gr (I. Iliopoulos).

https://doi.org/10.1016/j.csbj.2024.08.016
Received 19 April 2024; Received in revised form 5 August 2024; Accepted 15 August 2024
Available online 21 August 2024

meticulous tracking of advancements within specific domains and necessitates extracting actionable knowledge from scientific texts. This involves tasks such as identifying, visualizing, and analyzing connections between various biological entities and discovering novel concepts.

In the face of this data deluge, researchers increasingly rely on computational tools to sift through this information overload. Natural language processing methods, and especially text mining, have emerged as a vital tool to tackle the complexities inherent in biomedical research. Various applications have been developed to address diverse challenges, ranging from concept discovery in PubMed abstracts and OMIM records to identifying associations between entries in various databases.

Examples include BioTextQuest+ [1], which facilitates concept discovery in Pubmed abstracts and OMIM records, DrugQuest [2] for identifying associations between drugs and other entries of the DrugBank database, PolySearch [3] for identifying gene, mutation, and disease associations and DISEASES [4] for extracting disease-gene associations from biomedical abstracts. Tools like Darling [5] and OnTheFly [6] mine disease-related databases and perform named entity recognition in various texts and databases, respectively. UniProt Related Documents (UniReD) [7] assists wet lab biologists in their quest to find novel counterparts in a protein interaction network generated by text mining using UniProt records. CoPub is a text mining system for microarray data analysis but also for general exploration of biomedical literature [8]. Other tools, like EXTRACT [9], NETME [10], GNormPlus [11], and SciLite [12] detect associations between biomedical terms and construct knowledge networks incorporating term associations within the literature. PREGO [13] links environments to biological processes and organisms. PubAnnotation [14] is an open, Agile text mining framework assisting researchers for annotation purposes. In contrast, Medline Ranker [15] ranks abstracts from Medline, according to a training set of abstracts or a MeSH term. LipiDisease [16] focuses on identifying associations between lipids and diseases in biomedical literature data. PESCADOR [17] extracts a network of gene and protein interactions from a set of PubMed abstracts selected by a user and PubTator [18] offers automated annotations generated by cutting-edge text mining systems for genes/proteins, genetic variants, diseases, chemicals, species, and cell lines.

Databases such as STRING [19], STITCH [20]), DisGeNET [21], and CancerMine [22] periodically mine MEDLINE to computationally identify novel protein-protein, protein-chemical compound, and gene-disease associations to produce interaction networks, as well as enrich their existing evidence with additional annotations. Similarly, web servers offering NLP functionalities such as GePI[23]. Functional enrichment analysis tools such as aGOtool [24], SciMiner [25], ToppGene [26] and Flame [27] utilize text-mining annotations alongside other sources to annotate gene and protein lists. Finally, text-mining resources and datasets have been used to derive training sets for deep learning methods aimed at the functional characterization of unannotated proteins. Notable examples include ProtNLM which has been trained by text mining the annotation fields of UniProtKB [28] and Pfam [29] entries, and KV-PLM [30], trained using the S2ORC corpus of English academic research publications [31].

All the tools and resources mentioned above, along with numerous others, aid researchers in uncovering new knowledge within extensive lisis of biomedical articles, as well as in organizing and analyzing the vast amount of associated information.

Here, we present BioTextQuest v2.0 – an updated version of a previously published web tool BiotextQuest (+) – a tool for biomedical literature analysis and concept discovery. BioTextQuest v2.0 is an open-source online web portal for document clustering and efficient management of the overload of information in PubMed abstracts. It performs automated knowledge extraction and bridges the gap between named entity recognition and bioinformatics-related data sources. It takes as input PubMed queries and the output is presented in tabular and word cloud formats. In its current form, it utilizes EXTRACT, a tool that identifies genes/proteins, chemical compounds, organisms, environments, tissues, diseases, phenotypes, and Gene Ontology terms mentioned in Biomedical Literature. Substantial improvements of this significantly evolved version include: *i)* Use of Extract tool (BioTextQuest 2.0 recognizes more biomedical entities). *ii)* A user option to choose the type of biomedical entities to cluster. *iii)* Analysis of up to 10,000 PubMed abstracts (instead of 5000 in the previous version) *iv)* Implementation of two additional, state-of-the-art machine learning clustering algorithms. *v)* A more user-friendly interface, based also on suggestions from users, *vi)* Option to download the results in publication quality files/images and *vii*) Implementation of the Docker technology.

By enabling the user to choose the preferable biomedical entities to perform the clustering, users are assisted in their quest to identify all kinds of associations between these entities e.g. protein-disease associations, and thus extracting known or novel biomarkers. With all the enhancements, BioTextQuest v2.0 empowers researchers to navigate and extract valuable insights from the vast landscape of biomedical literature more effectively.

2. Methods

BioTextQuest version 2 is an easy-to-use web application with a responsive web interface. It is implemented in R [32] using the Shiny package[1] to build the web application. The Shiny package enables the creation of a user-friendly web interface aimed at enhancing the user experience. It provides a simple web interface with just a query field and graphical representation of clusters related to biological terms. To ensure high availability and scalability to internet-level usage without limitations on concurrent users of our Shiny app, we implemented the Shinyproxy open-source framework, which is built on JVM technology and Docker. The analysis process starts with entering a query into the designated field for the PubMed database. Users also have the option to customize analysis parameters through the advanced settings button. (Fig. 1).

The main difference from the previous version of BiotextQuest is that it uses extracted terms from the EXTRACT tagger [9]. EXTRACT is a text-mining-assisted interactive annotation app of biomedical named entities and ontology terms. It tags significant biomedical terms from PubMed abstracts, belonging to different categories, like gene ontology terms, diseases, genes, proteins, etc. The tagged terms are precomputed and stored locally in a MySQL database. BiotextQuest v.2 uses the tagged terms of each PubMed abstract of a query and clusters abstracts into subjects according to their similarity based on the extracted terms.

2.1. Query system

BioTextQuest v.2 currently queries PubMed. The PubMed database is locally stored in MongoDB. The query field allows input of any valid PubMed query supporting all features offered by the search mechanism of PubMed, such as field tags, Boolean operators, or grouping parentheses. BioTextQuest v.2 uses the easyPubMed R package to post a query directly to PubMed and obtain the PubMed identifiers of matching entries. In a subsequent step and depending on user-defined parameters, the platform uses these identifiers to retrieve the appropriate combination of abstracts and their associated EXTRACT tagged terms from the local database, thus maximizing the speed of information retrieval. The PubMed retrieval is performed in a few seconds. BioTextQuest v.2 users may also specify the number of documents to be retrieved/processed (with a maximum of 10,000 articles/entries as default). In cases of queries returning more than 10,000 abstracts, the 10,000 most recent are retained for analysis. The user can also choose specific categories of tagged terms, for example only genes/proteins to include in the analysis of abstracts.

[1] https://shiny.posit.co

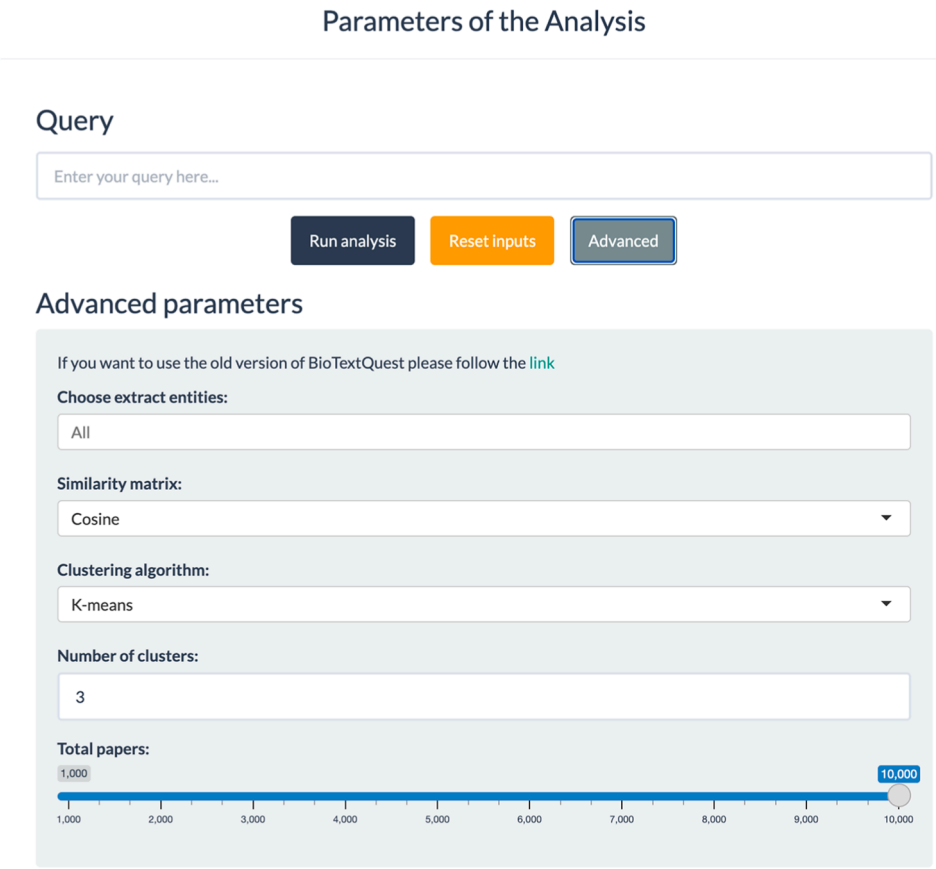

Fig. 1. The main page of BiotextQuest v.2 query and the Advanced parameters that users may define.

2.2. Document similarities

We represent each abstract with a binary vector indicating the presence or absence of the EXTRACT tagged terms found in the text collection. Similarity metrics available in BioTextQuest v.2 are the Cosine similarity, Jaccard, and Euclidean distance [33] that we have been deemed as best to obtain optimal results.

2.3. Document clustering

Based on our experience from the previous version of BiotextQuest, we incorporated four different clustering algorithms, namely K-means [34], MCL [35], Louvain [36] and Top2vec [37] that can be seen when pressing the Advanced parameters button at the web interface to cluster documents based on their EXTRACT tagged terms. The methods take as input the similarity matrices described in the document similarities section. In order not to overwhelm non-expert users, we chose to hide by default the relevant options from the main interface; K-means is selected as the default clustering algorithm with 3 clusters and the cosine similarity metric. Some may consider that the plethora of integrated algorithms can become confusing; however, it is a useful feature for experts, as each algorithm takes a different angle on how to cluster data. An important feature of BioTextQuest v.2 is that it hosts two classes of clustering algorithms, depending on whether the number of resulting clusters is required as input. For example, MCL automatically detects the number of clusters formed by the data (depending on the choice of inflation parameter), whereas K-Means requires that the number of clusters k is known beforehand. Empirically, the MCL and k-means clustering algorithms are among the fastest.

2.4. Visualization of results

The visualization of the results is mainly based on a tag cloud for each cluster. Each cluster's tag cloud contains by default the 50 most frequent EXTRACT tagged terms based on the abstracts grouped in that cluster. The terms are colored based on the category they belong to, for example, green is used for gene ontology terms, etc. The user can choose to include in each cluster more or less than 50 terms. Furthermore, the user can download the word clouds in publication quality figures as a PDF file. It is also noteworthy that the terms in each word cloud are links to the relevant external databases that contain information about the term, for example, GO terms are linked to the gene ontology database to retrieve all the relevant information. Users can also choose not to show tagged terms belonging to a specific category by deselecting it from the checkboxes at the right panel of the interface (Fig. 2).

Two additional tabs of results are available to assist users in delving deeper into the information generated by the analysis. The "Biomedical Terms" tab displays a comprehensive list of all the terms tagged with EXTRACT that have been retrieved, allowing users to sort them by frequency. Users can also search for specific terms within this tab (Fig. 3).

The second tab, named "Clustered Documents," provides an overview of the results, including the total number of clusters and the number of articles within each cluster. Additionally, this tab features a table that presents details about the abstracts utilized in the analysis, such as the PMID, abstract content, title, publication year, and the corresponding cluster assignment. Users have the option to filter and export the table based on their specific requirements. (Fig. 4).

3. Results and discussion

The use of BiotextQuest v.2 involves two primary steps. First, users commence their analysis by selecting the "Start" tab. This page hosts a

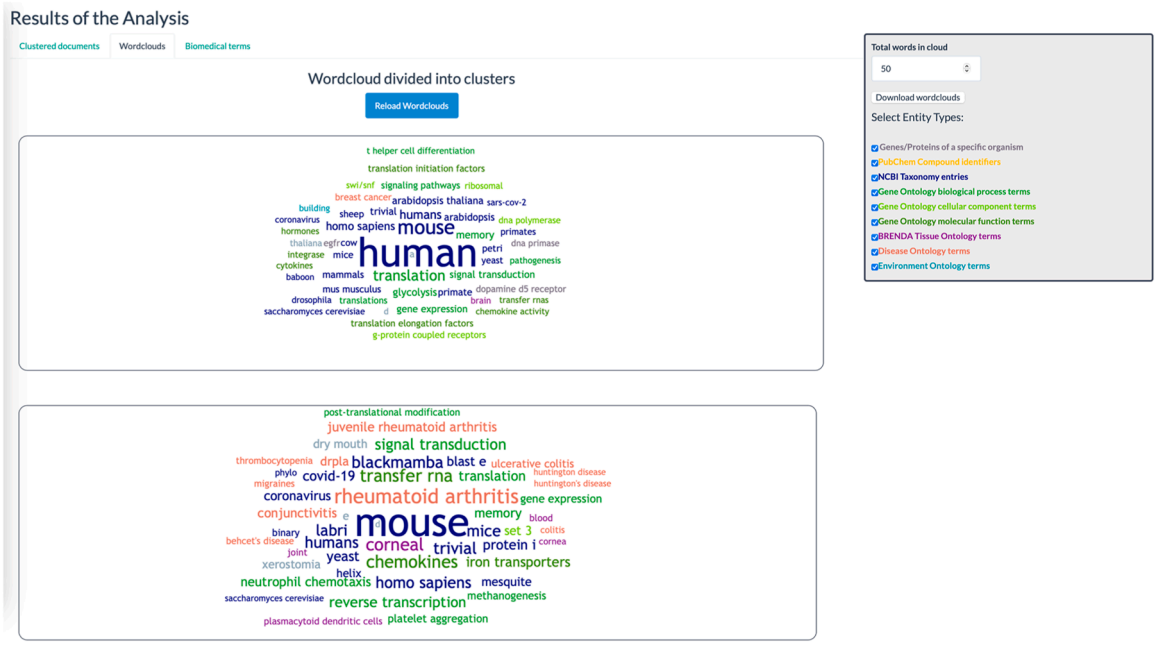

Fig. 2. Results presented as a word cloud.

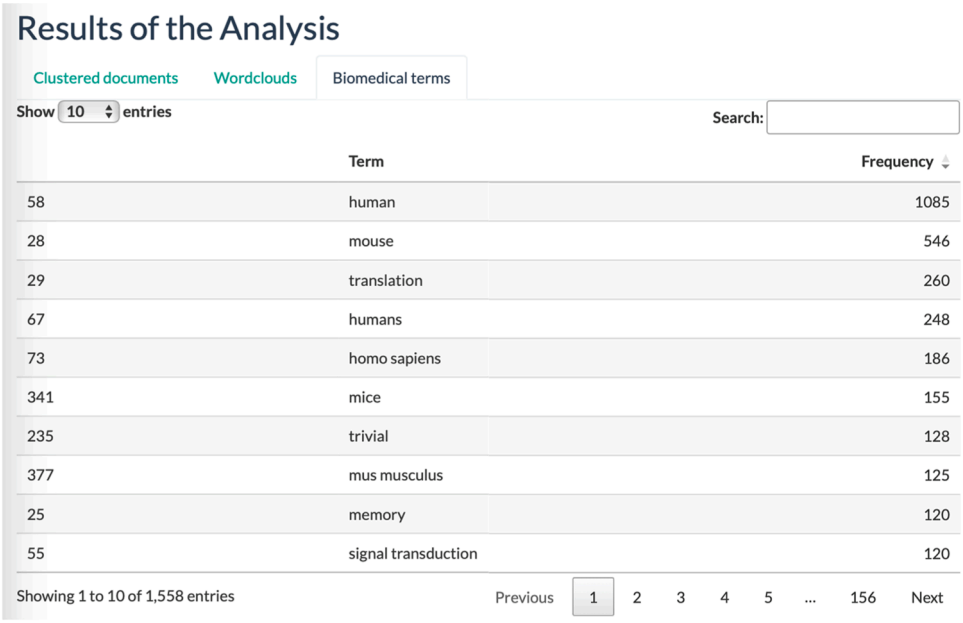

Fig. 3. The list of the tagged terms extracted from the abstracts and their frequency.

text field that allows input of any valid PubMed query, mirroring the syntax of PubMed's query system. Users also have the option to adjust analysis parameters via the "Advanced" button. The analysis is initiated by clicking the "Run Analysis" button, which triggers a progress bar display. Upon completion of the analysis, users are automatically directed to the "Results" tab. The initial page of results, labeled "Clustered documents" presents information such as the number of articles, their respective clusters, article titles, etc. By navigating to the "Wordclouds" tab, users gain access to interactive word clouds that showcase the EXTRACT tagged terms within each cluster. Finally, there is a third page of results called "Biological terms" containing in a sortable and filterable table all the EXTRACT terms utilized in the analysis of the PubMed articles.

Substantial improvements of this significantly evolved version include:

- Use of EXTRACT named entity recognition tool (BioTextQuest 2.0 recognizes more biomedical entities).
- The user has the option to choose the type of biomedical entities to cluster.
- Analysis of up to 10,000 PubMed abstracts (instead of 5000 in the previous version).
- Implementation of two additional and well-established clustering algorithms.
- A more user-friendly interface, based also on suggestions from users.
- Option to download the results in publication quality files/images.
- Implementation of Docker technology.

The usefulness of BioTextQuest v2.0 has been shown previously

Fig. 4. Information about the clustered documents.

elsewhere [38]. To assess the representation of the myeloid-derived suppressor cell (MDSC) concept and the tumor-associated neutrophils (TAN) concept in the literature, we employed the following text mining methodology using a preliminary version of Biotextquest version 2.0. We identified genes, proteins, molecular functions, pathways, and biological processes in PubMed abstracts. We used the following pattern to form our queries: "A and B and C," where A was "neutrophil" or "MDSC," B was always "cancer" and C was either "progression, suppression, mechanism, human, mice, maturation, clinic, or pathway." We analyzed sixteen queries in total and the advanced parameters were set as "use terms from the database: extract, choose extract entities: all, similarity matrix: cosine, clustering algorithm: K-means, number of clusters: 5.". Using the frequency of terms returned, as well as the word clouds of each cluster, we observed a clear focus of mentioning aspects of granulopoiesis and neutrophil differentiation (central term "G-CSF") in TAN-related research papers.

Key terms retrieved from MDSC literature reflected the central function of MDSC, which is their capacity to limit the activity of lymphocytes, in particular T cells (central words "CD4″ and "CD8″). The text mining also revealed the most prominent cell biological mechanism in the context of MDSC research, which appears to be linked to the transcription factor STAT3. This term appeared as a central word in connection with the search terms "progression", "suppression", "mechanism", "pathway" and "clinic". Our MDSC-TAN use case also revealed shared "key" words between MDSC and TAN, such as IL-6, which seems to be seen as relevant in both research areas.

Another use case involves RICTOR, a fundamental subunit of the mTORC2. The mechanistic target of rapamycin (mTOR) signaling pathway involves two distinct complexes, mTOR complex 1 (mTORC1) and 2 (mTORC2), which coordinate numerous vital cellular processes.

We used the following query to retrieve related abstracts from PubMed " Rictor ". This process retrieved 1143 abstracts and we analyze the articles using the default parameters. The result was three clusters and we set the number of terms in the wordcloud to be 100 (Fig. 1 supplementary data).

The mTORC2-RICTOR regulates numerous crucial cellular processes as indicated by the terms apoptosis, autophagy, cell growth, etc. (Wordcloud 2) [39], [40]. Genomic alterations in the mTOR pathway components are frequently detected in cancers, subsequently modifying pathway activity is implicated in oncogenesis of different tumor types [41]. There is growing evidence reporting that RICTOR is aberrantly regulated across numerous cancer types. Wordcloud 1 contains several types of cancers and more terms related to cancer than in the other two other Wordclouds, for example by the terms non-small cell lung cancer, ovarian cancer, melanoma, etc) and is associated with tumorigenesis and poor prognosis (Wordcloud 1) [42]. Moreover, different studies have underlined the correlation between RICTOR expression and glycolysis and other metabolic functions such as glycolysis and lipogenesis as indicated by the terms glycolysis, insulin receptor and fatty acid oxidation (Wordcloud 3) [43]. It must be noted that although some terms from the aforementioned examples are included in two clusters/wordclouds, their frequency is higher in the cluster they are more associated (table 1 supplementary data).

We then test the performance of BiotextQuest v.2 when compared to its predecessor BioTextQuest(+)[1] using the same case study. We performed four different MeSH-term–based queries and retrieved all PubMed articles that mention a specific phase of the human cell cycle by excluding all others (phases M, G1, S, G2). We used limitations in publication date in order to retrieve the same articles as in the previous test. The process returned 8 genes for M-phase, 7 for G1-Phase, 8 for S-phase and 11 for G2-phase, in the case of BioTextQuest(+) whereas using the updated version of BiotextQuest v.2 we were able to retrieve 60 genes for M-phase, 117 for G1-Phase, 138 for S-phase and 99 for G2-phase (see table 2 in supplementary data). This result clearly shows that BiotextQuest v.2 significantly outperforms the previous version.

In the future, BioTextQuest is expected to benefit from Large Language Models (LLMs) in text mining for biomedical literature. These advanced AI models could process vast amounts of scientific text with exceptional precision, revealing intricate relationships between genes, diseases, and drug targets. This capability will enable researchers to uncover hidden connections, generate novel hypotheses and accelerate the discovery process. The integration of LLMs into text mining has the potential to revolutionize biomedical research, leading to breakthroughs that could significantly advance human health. Moreover, by incorporating new types of biomedical text association studies, such as Publication-Wide Association Studies (PWAS), into BioTextQuest could open up new avenues for further understanding complex biological relationships. [44].

Even in its current form, BioTextQuest v2.0 proves to be a highly useful tool for analyzing extensive collections of articles. It supports biomedical researchers in uncovering new concepts and relationships between biological entities, including the discovery of biomarkers.

Summarizing, our tool can help researchers in prioritizing topics and focus on classical state-of-the art and systematic reviews on biomedical subjects in an unbiased manner. Since cited literature and chosen topics in such reviews are often and obviously subjective and mainly driven by the authors' individual knowledge, BioTextQuest v2.0 can help review authors identify central terms and themes, providing a balance between

individual study selection and systematic literature survey.

Source code and Docker image availability

The source code is available at: https://github.com/theodos/biotextquest_v2.

The docker image is available at: https://hub.docker.com/r/mpaltsai/biotextquest_4.2.2/tags.

Funding

This research has been supported by the project "ELIXIR-GR: Managing and Analysing Life Sciences Data" (MIS: 5002780), co-financed by Greece and the European Union—European Regional Development Fund. This work was supported by COST (European Cooperation in Science and Technology) Action 20117 - Converting molecular profiles of myeloid cells into biomarkers for inflammation and cancer (MyeInfoBank). This work has been supported by a research grant under the Horizon 2020 programme of the European Commission (TO_AITION, no. 848146). This work was also supported by Hellenic Foundation for Research and Innovation (H.F.R.I) under the call 'Greece 2.0 - Basic Research Financing Action (Horizontal support of all Sciences), Subaction II', Grant ID: 16718-PRPFOR; 'Greece 2.0 - National Recovery and Resilience Plan', Grant ID: TAEDR-0539180.

Author statement

We would like to submit the revised version of the article entitled: "BioTextQuest v2.0: an evolved tool for Biomedical Literature mining and concept discovery." by Theodosiou et. al. Below you can find a point-by-point description with answers to the reviewers' comments. We would like to thank you in advance for considering our article for publication and we are looking forward to hearing from you.

CRediT authorship contribution statement

Evangelos Andreakos: Writing – review & editing, Validation. **Sven Brandau:** Writing – review & editing, Validation. **Theodosios Theodosiou:** Writing – review & editing, Writing – original draft, Software, Methodology. **Ekaterini Chatzaki:** Writing – review & editing, Validation. **Konstantinos Vrettos:** Software. **Christos A. Ouzounis:** Writing – review & editing, Validation. **Dimitrios Mossialos:** Validation. **Makrina Karaglani:** Validation. **Vasilis J. Promponas:** Writing – review & editing, Validation. **Fotis Baltoumas:** Software. **Ioannis Iliopoulos:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Ismini Baltsavia:** Software. **Andreas N. Antonakis:** Validation. **Nikolas Papanikolaou:** Validation. **Georgios A. Pavlopoulos:** Writing – review & editing, Writing – original draft, Validation.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

We would like to thank Anastasios Gkountakos for his assistance in the "Rictor" use case.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.08.016.

References

[1] Papanikolaou N, et al. BioTextQuest(+): a knowledge integration platform for literature mining and concept discovery. Bioinformatics 2014;vol. 30(22): 3249–56. https://doi.org/10.1093/bioinformatics/btu524.
[2] Papanikolaou N, Pavlopoulos GA, Theodosiou T, Vizirianakis IS, Iliopoulos I. DrugQuest - a text mining workflow for drug association discovery. BMC Bioinforma 2016;vol. 17(Suppl 5):182. https://doi.org/10.1186/s12859-016-1041-6.
[3] Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 2008;vol. 36 (Web Server):W399–405. https://doi.org/10.1093/nar/gkn296.
[4] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods 2015;vol. 74: 83–9. https://doi.org/10.1016/j.ymeth.2014.11.020.
[5] Karatzas E, et al. Darling: a web application for detecting disease-related biomedical entity associations with literature mining. Biomolecules 2022;vol. 12 (4). https://doi.org/10.3390/biom12040520.
[6] Baltoumas FA, et al. OnTheFly2.0: a text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis. NAR Genom Bioinform 2021;vol. 3(4):lqab090. https://doi.org/10.1093/nargab/lqab090.
[7] Theodosiou T, et al. UniProt-Related Documents (UniReD): assisting wet lab biologists in their quest on finding novel counterparts in a protein network. NAR Genom Bioinform 2020;vol. 2(1):lqaa005. https://doi.org/10.1093/nargab/lqaa005.
[8] Fleuren WWM, et al. CoPub update: CoPub 5.0 a text mining system to answer biological questions. Nucleic Acids Res 2011;vol. 39(Web Server issue):W450–4. https://doi.org/10.1093/nar/gkr310.
[9] Pafilis E, et al. EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. Database (Oxf) 2016;vol. 2016. https://doi.org/10.1093/database/baw005.
[10] Muscolino A, et al. NETME: on-the-fly knowledge network construction from biomedical literature. Appl Netw Sci 2022;vol. 7(1):1. https://doi.org/10.1007/s41109-021-00435-x.
[11] Wei C-H, Kao H-Y, Lu Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. Biomed Res Int 2015;vol. 2015:1–7. https://doi.org/10.1155/2015/918710.
[12] Venkatesan A, et al. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. Wellcome Open Res 2017;vol. 1:25. https://doi.org/10.12688/wellcomeopenres.10210.2.
[13] Zafeiropoulos H, Paragkamian S, Ninidakis S, Pavlopoulos GA, Jensen LJ, Pafilis E. PREGO: a literature and data-mining resource to associate microorganisms, biological processes, and environment types. Microorganisms 2022;vol. 10(2):293. https://doi.org/10.3390/microorganisms10020293.
[14] Kim J-D, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 2019;vol. 35(21):4372–80. https://doi.org/10.1093/bioinformatics/btz227.
[15] Fontaine J-F, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA. MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Res 2009;vol. 37(suppl_2):W141–6. https://doi.org/10.1093/nar/gkp353.
[16] More P, Bindila L, Wild P, Andrade-Navarro M, Fontaine J-F. LipiDisease: associate lipids to diseases using literature mining. Bioinformatics 2021;vol. 37(21):3981–2. https://doi.org/10.1093/bioinformatics/btab559.
[17] Barbosa-Silva A, Fontaine J-F, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinforma 2011;vol. 12(1):435. https://doi.org/10.1186/1471-2105-12-435.
[18] Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res 2019;vol. 47(W1):W587–93. https://doi.org/10.1093/nar/gkz389.
[19] Szklarczyk D, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021;vol. 49(D1):D605–12. https://doi.org/10.1093/nar/gkaa1074.
[20] Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res 2016;vol. 44(D1):D380–4. https://doi.org/10.1093/nar/gkv1277.
[21] Piñero J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2019. https://doi.org/10.1093/nar/gkz1021.
[22] Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods 2019;vol. 16(6):505–7. https://doi.org/10.1038/s41592-019-0422-y.
[23] Faessler E, Hahn U, Schäuble S. GEPI: large-scale text mining, customized retrieval and flexible filtering of gene/protein interactions. Nucleic Acids Res 2023;vol. 51 (W1):W237–42. https://doi.org/10.1093/nar/gkad445.
[24] Schölz C, Lyon D, Refsgaard JC, Jensen LJ, Choudhary C, Weinert BT. Avoiding abundance bias in the functional annotation of posttranslationally modified proteins. Nat Methods 2015;vol. 12(11):1003–4. https://doi.org/10.1038/nmeth.3621.
[25] Hur J, Schuyler AD, States DJ, Feldman EL. SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics Mar. 2009;vol. 25(6):838–40. https://doi.org/10.1093/bioinformatics/btp049.

[26] Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 2009;vol. 37(Web Server):W305–11. https://doi.org/10.1093/nar/gkp427.
[27] Karatzas E, et al. Flame (v2.0): advanced integration and interpretation of functional enrichment results from multiple sources. Bioinformatics 2023;vol. 39 (8). https://doi.org/10.1093/bioinformatics/btad490.
[28] Bateman A, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 2023;vol. 51(D1):D523–31. https://doi.org/10.1093/nar/gkac1052.
[29] Mistry J, et al. Pfam: The protein families database in 2021. Nucleic Acids Res 2021;vol. 49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913.
[30] Zeng Z, Yao Y, Liu Z, Sun M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. Nat Commun 2022;vol. 13(1):862. https://doi.org/10.1038/s41467-022-28494-3.
[31] K. Lo, L.L. Wang, M. Neumann, R. Kinney, and D. Weld, "S2ORC: The Semantic Scholar Open Research Corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 4969–4983. doi: 10.18653/v1/2020.acl-main.447.
[32] R Core Team, "R: A Language and Environment for Statistical Computing," 2022, *Vienna, Austria*: https://www.R-project.org/.
[33] Singh Lehal Manpreet. Comparison of Cosine, Euclidean Distance and Jaccard Distance. Int J Sci Res Sci, Eng Technol(IJSRSET) 2017;vol. 3(8):1376–81.
[34] Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;vol. 28(2): 129–37.
[35] Van Dongen S. Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl 2008;vol. 30(1):121–41. https://doi.org/10.1137/040608635.
[36] Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech: Theory Exp Oct. 2008;vol. 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.
[37] D. Angelov, "Top2Vec: Distributed Representations of Topics," *ArXiv*, vol. abs/2008.09470, 2020.
[38] Antuamwine BB, et al. N1 versus N2 and PMN-MDSC: a critical appraisal of current concepts on tumor-associated neutrophils and new directions for human oncology. Immunol Rev 2023;vol. 314(1):250–79. https://doi.org/10.1111/imr.13176.
[39] Lee G, Chung J. Discrete functions of rictor and raptor in cell growth regulation in Drosophila. Biochem Biophys Res Commun 2007;vol. 357(4):1154–9. https://doi.org/10.1016/j.bbrc.2007.04.086.
[40] Ballesteros-Álvarez J, Andersen JK. mTORC2: The other mTOR in autophagy regulation. Aging Cell Aug. 2021;vol. 20(8). https://doi.org/10.1111/acel.13431.
[41] Saxton RA, Sabatini DM. mTOR signaling in growth, metabolism, and disease. Cell 2017;vol. 168(6):960–76. https://doi.org/10.1016/j.cell.2017.02.004.
[42] Gkountakos A, et al. Unmasking the impact of Rictor in cancer: novel insights of mTORC2 complex. Carcinogenesis 2018;vol. 39(8):971–80. https://doi.org/10.1093/carcin/bgy086.
[43] Kocalis HE, et al. Rictor/mTORC2 facilitates central regulation of energy and glucose homeostasis. Mol Metab 2014;vol. 3(4):394–407. https://doi.org/10.1016/j.molmet.2014.01.014.
[44] Narganes-Carlón David, Crowther Daniel J, Pearson Ewan R. A publication-wide association study (PWAS), historical language models to prioritise novel therapeutic drug targets. Sci Rep 2023;vol. 13(1):8366. https://doi.org/10.1038/s41598-023-35597-4.