Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210124

COVID-19 preVIEW: Semantic Search to Explore COVID-19 Research Preprints

Lisa LANGNICKEL[a,b], Roman BAUM[a], Johannes DARMS[a], Sumit MADAN[c,d] and Juliane FLUCK[a,c,d,1]
[a]*ZB MED - Information Centre for Life Sciences, Cologne, Germany*
[b]*Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Faculty of Technology, Bielefeld University, Bielefeld, Germany*
[c]*University of Bonn, Bonn, Germany*
[d]*Fraunhofer Institute for Algorithms and Scientific Computing, St. Augustin, Germany*

Abstract. During the current COVID-19 pandemic, the rapid availability of profound information is crucial in order to derive information about diagnosis, disease trajectory, treatment or to adapt the rules of conduct in public. The increased importance of preprints for COVID-19 research initiated the design of the preprint search engine preVIEW. Conceptually, it is a lightweight semantic search engine focusing on easy inclusion of specialized COVID-19 textual collections and provides a user friendly web interface for semantic information retrieval. In order to support semantic search functionality, we integrated a text mining workflow for indexing with relevant terminologies. Currently, diseases, human genes and SARS-CoV-2 proteins are annotated, and more will be added in future. The system integrates collections from several different preprint servers that are used in the biomedical domain to publish non-peer-reviewed work, thereby enabling one central access point for the users. In addition, our service offers facet searching, export functionality and an API access. COVID-19 preVIEW is publicly available at https://preview.zbmed.de.

Keywords. COVID-19, Information Retrieval, Biomedical Text Mining

1. Introduction

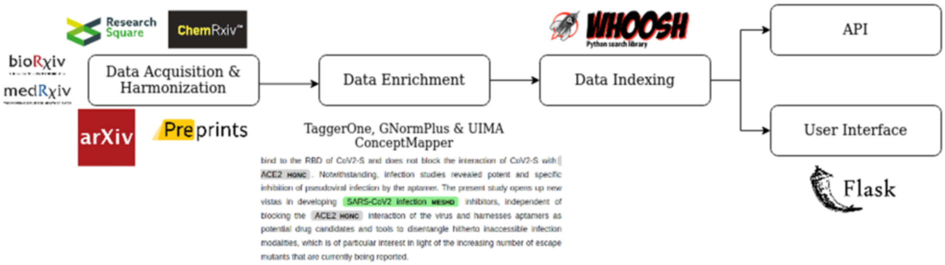
To find relevant literature, traditional search engines such as PubMed [1], Europe PMC [2] or LIVIVO [3] are commonly used in the biomedical domain. Since the beginning of the COVID-19 pandemic, publication of results prior to the usual review process became even more necessary due to an urgent need for information and knowledge exchange. As the peer-review process can take up to several months for a single publication, preprint servers allow faster access and gained popularity. An increasing number of datasets and services provide access to both peer-reviewed journal articles and preprints. The COVID-19 Open Research Dataset (CORD-19) [4] integrates various publications in raw form. Europe PMC services provide web-based access to several preprints since 2018 and PubMed started a pilot project to integrate COVID-19 related preprints into their database [5]. Based on the user demands of German epidemiologists to differentiate between reviewed publications and preprints, we set up

[1] Corresponding Author, Juliane Fluck, ZB MED - Information Centre for Life Sciences; E-mail: fluck@zbmed.de

preVIEW, a user-friendly COVID19-related preprint viewer with advanced semantic search functionality. preVIEW currently enables central access to six different preprint servers and supports search functionality and corpus analyses through automatic enrichment with biomedical annotations for diseases, human genes and SARS-CoV-2 proteins. Moreover, we provide an easy-to-use API to access data programmatically.

2. Methods

COVID-19 preVIEW aggregates several preprint data sources where each source provides a different data format. Therefore, as a first step, a data acquisition and harmonization component is responsible for data retrieval and conversion into a common data model. The second component enriches this data with biomedical annotations based on an integrated named entity recognition (NER) process. Finally, two end-user components, a web-based user interface (WebUI) and an API, expose the semantic search functionalities to the user. Figure 1 depicts the key components and shows how data is passed between the components.


Research Square
ChemRxiv
bioRxiv
medRxiv
arXiv
Preprints
Data Acquisition & Harmonization
Data Enrichment
TaggerOne, GNormPlus & UIMA ConceptMapper
WHOOSH Python search library
Data Indexing
API
User Interface
Flask
ACE2 HGNC
SARS-CoV2 infection MESHD

bind to the RBD of CoV2-S and does not block the interaction of CoV2-S with ACE2 HGNC . Notwithstanding, infection studies revealed potent and specific inhibition of pseudoviral infection by the aptamer. The present study opens up new vistas in developing SARS-CoV2 infection MESHD inhibitors, independent of blocking the ACE2 HGNC interaction of the virus and harnesses aptamers as potential drug candidates and tools to disentangle hitherto inaccessible infection modalities, which is of particular interest in light of the increasing number of escape mutants that are currently being reported.

Figure 1. The key components of the COVID-19 preVIEW system embedded in a workflow.

For all preprint servers, we are using specific API calls to retrieve the corresponding publication metadata. Until now, six different preprint servers (bioRxiv, medRxiv, aRxiv, ResearchSquare, ChemRxiv and Preprints.org) have been integrated. Whereas bioRxiv and medRxiv provide a COVID-19 specific website with the opportunity to access JSON data files directly, specific search queries were defined for all other resources. The retrieved data is further transformed and embedded into a common data model, which covers the DOI (if available), the authors, the source (i.e. preprint server), the title, the abstract and the date of publication.

Enriching the data with semantic concepts or relations is key for building a semantic search. In the current setting of preVIEW, diseases, genes and SARS-CoV-2 proteins are annotated by an NER workflow. We use TaggerOne [6], a joint named entity recognition and normalization tool, to detect MeSH disease terms. To annotate human genes and proteins, we integrated GNormPlus [7] and link the concepts to HUGO Gene Nomenclature Committee (HGNC) database identifiers. In addition, we use the annotation engine ConceptMapper [8] with our manually developed SARS-CoV-2 protein terminology to detect SARS-CoV-2 protein names, which we provide under https://github.com/zbmed/preVIEW-COVID-19 (as well as further method descriptions). To enable semantic search functionality, we index metadata of all documents including the biomedical annotations obtained through text mining with the Python library Whoosh [9]. The system offers the possibility to search each field and allows building more complex search queries by combining sub queries using Boolean operators. Furthermore, semantic queries are enabled by indexing the identifiers of the biomedical concepts.

3. Results

3.1. Data and NER

The data acquisition results in more than 24,600 COVID-19 related articles that can be queried (05/03/2021). Most of the articles originate from medRxiv (43.1%), 23.8% are retrieved from ResearchSquare, followed by 12.6% coming from bioRxiv. For evaluation purpose, a total of 100 documents have been manually annotated with the annotation tool Doccano [10]. The results of inter-annotator agreement (IAA) and NER recognition are summarized in Table 1. For all three entity classes, we have high IAA between 93.95% and 97.87% F1-score for relaxed matches (ignoring one text boundary) and between 87.77% and 91.10% F1-score for exact matches. The automatic NER performs very well for diseases and SARS-CoV-2 proteins with F1-scores of 90.16% and 94.42% for relaxed matches. However, for human gene recognition only an F1-score of 74% was reached due to a low precision of 61.58%. The error analysis revealed a number of false matches annotating virus proteins as human proteins.

Table 1. Inter-annotator agreement and NER results

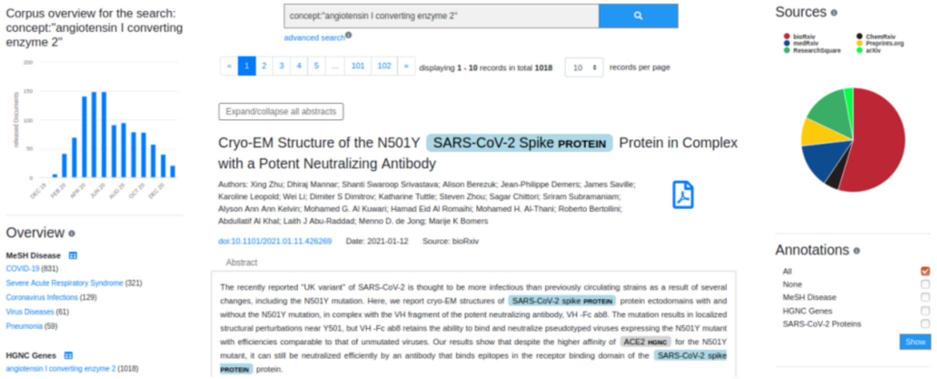
	Entity Class	Precision [%]	Recall [%]	F1-Score [%]
IAA	Disease	96.35 (93.43)	91.67 (88.89)	93.95 (91.10)
Modified TaggerOne		85.93 (80.93)	94.82 (89.31)	90.16 (84.92)
IAA	Gene	98.57 (89.70)	97.18 (85.92)	97.87 (87.77)
GNorm Plus		61.58 (59.80)	92.72 (90.04)	74.00 (71.87)
IAA	SARS-CoV-2 proteins	95.36 (90.07)	93.51 (88.31)	94.42 (89.18)
ConceptMapper		90.00 (75.78)	83.41 (70.24)	86.58 (72.91)

3.2. WebUI and API

The COVID-19 preVIEW WebUI shows users an overview of the most recent preprint publications. For each publication metadata such as title, authors, DOI (if available), date of publication, and source for each publication are shown. Furthermore, direct links to both the full-text PDF version as well as to the article on the corresponding preprint server are given. Detected concepts are highlighted, additional information (such as synonyms, concept definition, ontology hierarchy) can be seen via a mouse-hover or clicking on the term. To find relevant literature, a search bar is provided on top that enables both simple and complex search queries. Additional facets, provided on the left side, allow the users to further filter the search results. A screenshot of the COVID-19 preVIEW system is depicted in Figure 2. The first facet allows users to filter results based on their publishing dates in monthly intervals. Second facet shows the most abundant concepts for diseases, genes and SARS-CoV-2 proteins. Furthermore, a table icon guides the user to a detailed overview of all detected concepts with their frequencies. A selection of a facet item refines the search.

On the right hand side, the data distribution of the preprint servers is visualized for the current search query. Furthermore, the user can filter the search result by selecting the resources. The user can also customize the visualization of concepts by enabling or disabling the highlighting of certain annotations. An export of the results either in Endnote or BibTeX bibliography formats is available.

The RESTful API with documentation and examples is available at https://preview.zbmed.de/api.



Overview

MeSH Disease

COVID-19 (831)

Severe Acute Respiratory Syndrome (321)

Coronavirus Infections (129)

Virus Diseases (61)

Pneumonia (59)

HGNC Genes

angiotensin I converting enzyme 2 (1018)

concept:"angiotensin I converting enzyme 2"

displaying **1 - 10** records in total **1018** 10 records per page

Cryo-EM Structure of the N501Y SARS-CoV-2 Spike PROTEIN Protein in Complex with a Potent Neutralizing Antibody

Authors: Xing Zhu; Dhiraj Mannar; Shanti Swaroop Srivastava; Alison Berezuk; Jean-Philippe Demers; James Saville; Karoline Leopold; Wei Li; Dimiter S Dimitrov; Katharine Tuttle; Steven Zhou; Sagar Chittori; Sriram Subramaniam; Alyson Ann Ann Kelvin; Mohamed G. Al Kuwari; Hamad Eid Al Romaihi; Mohamed H. Al-Thani; Roberto Bertollini; Abdullatif Al Khal; Laith J Abu-Raddad; Menno D. de Jong; Marije K Bomers

doi:10.1101/2021.01.11.426269 Date: 2021-01-12 Source: bioRxiv

Abstract

The recently reported "UK variant" of SARS-CoV-2 is thought to be more infectious than previously circulating strains as a result of several changes, including the N501Y mutation. Here, we report cryo-EM structures of SARS-CoV-2 spike PROTEIN protein ectodomains with and without the N501Y mutation, in complex with the VH fragment of the potent neutralizing antibody, VH -Fc ab8. The mutation results in localized structural perturbations near Y501, but VH -Fc ab8 retains the ability to bind and neutralize pseudotyped viruses expressing the N501Y mutant with efficiencies comparable to that of unmutated viruses. Our results show that despite the higher affinity of ACE2 HGNC for the N501Y mutant, it can still be neutralized efficiently by an antibody that binds epitopes in the receptor binding domain of the SARS-CoV-2 spike PROTEIN protein.

Annotations

All ☑

None □

MeSH Disease □

HGNC Genes □

SARS-CoV-2 Proteins □

Figure 2. A screenshot of the COVID-19 preVIEW system.

4. Discussion

Especially in times of a pandemic, the rapid generation and sharing of new knowledge is indispensable. Hence, the amount of available preliminary research results in the form of preprints is increasing enormously and new centralized access modes are necessary to integrate the relevant preprint services. Although several search engines such as PubMed and Europe PMC now include preprints in their COVID-19 search set, we believe it helps the research community by providing a separate service to distinct between peer reviewed articles and preprints.

The presented semantic search engine COVID-19 preVIEW provides a single entry point to search for COVID-19 related preprints aggregating six different preprint servers. In future, we intend to extend our service with further resources.

In the current version of our service, we provide a user interface with a high number of functionalities such as graphical overviews with simple querying of publications for certain time periods or preprint sources and provide also advanced filter opportunities.

Furthermore, through an automatic NER workflow, we enable semantic search. Other prominent examples of semantic search engines are PubTator Central [11] and SCAIView [12], which both offer biomedical annotations such as diseases, genes or drugs and statistical overviews of the most frequently annotated terms. Similarly, but restricted to full text, Europe PMC offers text mining based annotations. In preprints, automatic indexing is especially important since no other indexing has been addressed to this category. Although these annotations are not always correct, we could show annotation performance over 90% F1-score for two out of three classes. COVID-19 preVIEW has been developed based on the acute need during the current pandemic. However, the simple, modular design makes extensions to other entity classes or usage for new use cases easily possible.

5. Conclusions

In this work, we presented a COVID-19 related preprint viewer which, to the best of our knowledge, is the first semantic search service that allows unified access to resources published in different preprint servers. It provides both web-based and programmatic access, and enriches publications from six different preprint servers with biomedical annotations. While this service has been developed relating to the current COVID-19 pandemic, the easy and modular system allows expansion to different use cases.

Acknowledgement

This work was done as part of the NFDI4Health Task Force COVID-19. We gratefully acknowledge the financial support of the German Research Foundation (DFG) – Project Number 451265285. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). SM was supported by the MAVO program of the Fraunhofer Society. Moreover, we would like to thank Dr. Theresa Idda (ZB MED Information Centre for Life Sciences), Dr. Berit Lange (Helmholtz Centre for Infection Research), Dr. Holger Kirsten (University of Leipzig, Institute for Medical Informatics, Statistics and Epidemiology) and Jessica Sturm (MF 4/Library, Robert Koch Institute) for their contribution to corpus annotation, the feature specification, selection of preprint servers and feedback on the usability. We also would like to thank Dr. Klaus Lippert (ZB MED, Information Centre for Life Sciences) for the support in software development.

References

[1] PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/, Accessed March 3,2021.

[2] Levchenko M, Gou Y, Graef F, Hamelers A, Huang Z, Ide-Smith M, et al. Europe PMC in 2017. Nucleic Acids Res. 2018 04;46(D1):D1254–60.

[3] Pössel J. LIVIVO: Das neue ZB MED-Suchportal Lebenswissenschaften. GMS Med Bibl Inf. 2015 Dec 21;15(3):Doc25.

[4] Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. CORD-19: The COVID-19 Open Research Dataset. 2020 Jul 10; Available at: http://arxiv.org/abs/2004.10706, Accessed March 3,2021.

[5] NIH Preprint Pilot. Available at: https://www.ncbi.nlm.nih.gov/pmc/about/nihpreprints, Accessed March 3,2021.

[6] Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics. 2016 Sep 15;32(18):2839–46.

[7] Wei C-H, Kao H-Y, Lu Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains, BioMed Research International 2015.

[8] Funk C, Baumgartner W, Garcia B, Roeder C, Bada M, Cohen KB, et al. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinformatics. 2014 Feb 26;15(1):59.

[9] Introduction to Whoosh — Whoosh 2.7.4 documentation. Available at: https://whoosh.readthedocs.io/en/latest/intro.html, Accessed January 14,2021.

[10] Nakayama H, Kubo T, Taniguchi Y, Liang X. doccano: Text Annotation Tool for Human. doccano; 2021. Available at: https://github.com/doccano/doccano, Accessed January 14,2021.

[11] Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019 Jul 2;47(W1):W587–93.

[12] Younesi E, Toldo L, Müller B, Friedrich CM, Novac N, Scheer A, et al. Mining biomarker information in biomedical literature. BMC Med Inform Decis Mak. 2012 Dec 18;12:148.