



Databases and ontologies

COVID-19 Knowledge Graph from semantic integration of biomedical literature and databases

Chuming Chen ^{1,*}, Karen E. Ross², Sachin Gavali ¹, Julie E. Cowart¹ and Cathy H. Wu^{1,2}

¹Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA and ²Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC 20007, USA

*To whom correspondence should be addressed.

Associate Editor: Zhiyong Lu

Received on May 2, 2021; revised on September 26, 2021; editorial decision on October 2, 2021; accepted on October 4, 2021

Abstract

Summary: The global response to the COVID-19 pandemic has led to a rapid increase of scientific literature on this deadly disease. Extracting knowledge from biomedical literature and integrating it with relevant information from curated biological databases is essential to gain insight into COVID-19 etiology, diagnosis and treatment. We used Semantic Web technology RDF to integrate COVID-19 knowledge mined from literature by iTextMine, PubTator and SemRep with relevant biological databases and formalized the knowledge in a standardized and computable COVID-19 Knowledge Graph (KG). We published the COVID-19 KG via a SPARQL endpoint to support federated queries on the Semantic Web and developed a knowledge portal with browsing and searching interfaces. We also developed a RESTful API to support programmatic access and provided RDF dumps for download.

Availability and implementation: The COVID-19 Knowledge Graph is publicly available under CC-BY 4.0 license at <https://research.bioinformatics.udel.edu/covid19kg/>.

Contact: chenc@udel.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The worldwide research community's response to the COVID-19 pandemic has led to a burst of publications on this deadly disease (Brainard, 2020). The need for computational approaches and tools that can distill biomedical knowledge from literature and integrate it with relevant information from curated biological databases is essential to gain insight into COVID-19 etiology, diagnosis and treatment. Chen *et al.* (2021a) has surveyed more than 200 natural language processing studies and systems addressing the COVID-19 pandemic. Knowledge Graphs (KGs) are a powerful method to represent and integrate such heterogeneous data and their relationships to generate novel insights. Several efforts are underway to investigate COVID-19 using KGs. A cause-and-effect KG on COVID-19 pathophysiology was constructed from literature (Domingo-Fernandez *et al.*, 2021). A framework that can integrate heterogeneous biomedical data to produce KGs was developed for COVID-19 (Reese *et al.*, 2021). Repurposing drugs were discovered using a literature-derived KG and the graph completion method (Zhang *et al.*, 2021). A detailed review comparing existing KGs and our work can be found in [Supplementary File S1](#).

In this article, we used Semantic Web technology RDF (Resource Description Framework) to integrate COVID-19 knowledge from literature annotated by text-mining pipelines as well as relevant

biological databases. Information was extracted and formalized in a standardized and computable KG to enable researchers to explore, analyze and answer questions. To make this resource readily available to the research community in accordance with the FAIR principles (Wilkinson *et al.*, 2016), we published the COVID-19 KG with multiple dissemination mechanisms, including a SPARQL (RDF Query Language) endpoint, a knowledge portal, a RESTful API, as well as downloadable RDF dumps.

2 Materials and methods

LitCovid is a curated resource of articles about COVID-19 and SARS-CoV-2 in PubMed (Chen *et al.*, 2021b). The COVID-19 Open Research Dataset (CORD-19) (Wang *et al.*, 2020) consists of publications and preprints on COVID-19 and other coronaviruses (SARS and MERS) from the WHO, PubMed Central, bioRxiv and medRxiv. The abstracts and full texts from LitCovid and CORD-19 datasets have been processed by several text-mining pipelines to discover entity and relationship annotations: (i) iTextMine (Ren *et al.*, 2018), which provides text mining relation extraction results for protein phosphorylation (kinase-substrate-site), phosphorylation-dependent protein-protein interactions and miRNA-gene relations; (ii) PubTator (Wei *et al.*, 2019), which provides annotations of

biomedical concepts such as genes/proteins, genetic variants, diseases, chemicals, species and cell lines; and (iii) SemRep (Roseblat et al., 2013), which uses the Unified Medical Language System (Humphreys et al., 1998) to extract semantic predictions from biomedical text. We also include relevant data from curated biomedical databases such as Protein Ontology (Chen et al., 2020), DrugBank (Wishart et al., 2018), CoV-AbDab (Raybould et al., 2021), UniProtKB (UniProt Consortium, 2020), STRING (Szklarczyk et al., 2019) and iPTMnet (Huang et al., 2018).

The annotations of the LitCovid and COVID-19 datasets by iTextMine and PubTator in BioC JSON format were downloaded and converted to RDF format using AtomGraph's generic JSON to RDF converter. DrugBank data in XML format was downloaded and converted to JSON using xml2json tool, then converted to RDF format. CoV-AbDab, SemRep, STRING and iPTMnet data in text format were downloaded and converted to RDF format using custom scripts. The source code and instructions on how to create those RDF files used in the COVID-19 KG are publicly available at https://github.com/udel-cbcb/covid19kg_rdf.

The COVID-19 KG is served by OpenLink Virtuoso server community edition with SPARQL 1.1 query federation. To help the exploration and use of COVID-19 KG, a knowledge portal (<https://research.bioinformatics.udel.edu/covid19kg/>) with browsing and searching interfaces was developed using Django framework. The KG can be accessed via YASGUI with comprehensive example SPARQL queries for new users. We also developed a RESTful API for programmatic access to KG for data integration and analysis. In addition, we provide RDF dumps of COVID-19 KG in text/turtle format with corresponding RDF centric statistics.

3 Results and future work

The COVID-19 KG consists of 23 Named Graphs with a total of more than 1.2 billion RDF triples. The summary statistics of literature sources and the entities and relationships annotated by different text-mining tools can be found at the knowledge portal under 'Dashboard'.

For case studies, we have used the COVID-19 KG to identify drug repurposing candidates for COVID-19 and potential therapeutic interventions to disrupt function of the SARS coronavirus nucleocapsid protein (N protein). Detailed descriptions can be found in [Supplementary File S2](#).

To construct a drug repurposing network, we used the COVID-19 KG SPARQL GUI (query CPPQ5) to retrieve the top 10 most frequently mentioned genes in the COVID-19 corpus as annotated by PubTator. We then browsed the DrugBank section of the KG web interface to identify drug and disease relations involving these genes. Finally, we performed a federated SPARQL query with DisGeNET (Piñero et al., 2020) and a web search of the Therapeutic Information Browser (TIB) (https://covidtib.c19hcc.org/app_direct/dashboard/) for additional variant, drug and disease relations. In August 2020, our network predicted that the TNF-targeting drugs *etanercept* and *certolizumab pegol* ([Supplementary Fig. S1A](#)) were candidates for COVID-19 drug repurposing. As of March 2021, both drugs were mentioned in the COVID-19 literature and *etanercept* has been reported to be beneficial in individual cases (Clark, 2020; Zhu et al., 2021). Another promising candidate identified using the KG, the IFN- γ targeting drug, *olsalazine* ([Supplementary Fig. S1B](#)), is currently not mentioned in the COVID-19 literature. However, other IFN- γ targeting drugs are being investigated as COVID-19 treatments. Moreover, *olsalazine* is a recommended treatment for ulcerative colitis (UC), and several other UC therapeutics are mentioned in the literature in the context of COVID-19.

To identify strategies to disable the N protein ([Supplementary Fig. S2](#)), we browsed the KG using the web interface to identify phosphorylation and protein-protein interaction relations involving the N protein. We then further browsed the KG to identify drugs and miRNAs that targeted kinases that phosphorylate N protein

and proteins that interact with it. The potential avenues of intervention we identified include small molecule inhibitors of the N protein kinases, CDK1 and GSK3 ([Supplementary Fig. S2](#), V-shaped nodes) or miRNAs that inhibit expression of LARP1 and/or G3BP1 ([Supplementary Fig. S2](#), triangular nodes).

The COVID-19 KG will be regularly updated. We plan to develop a visualization application for the KG and combine graph representation learning, ontology, automated reasoning and neural networks to open up the KG for machine learning and further data analytics.

Funding

This work was partially supported by the National Institutes of Health [U24HG007822 and R35GM141873] and institutional resources at the University of Delaware.

Conflict of Interest: none declared.

References

- Brainard, J. (2020) Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science*, doi: 10.1126/science.abc7839.
- Chen, C. et al. (2020) Protein ontology on the semantic web for knowledge discovery. *Sci. Data*, 7, 337.
- Chen, Q. et al. (2021a) Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing. *Annu. Rev. Biomed. Data Sci.*, 4, 313–339.
- Chen, Q. et al. (2021b) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, 49, D1534–D1540.
- Clark, I. (2020) Background to new treatments for COVID-19, including its chronicity, through altering elements of the cytokine storm. *Rev. Med. Virol.*, 31, 1–31.
- Domingo-Fernández, D. et al. (2021) COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics*, 37, 1332–1334.
- Huang, H. et al. (2018) iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Res.*, 46, D542–D550.
- Humphreys, B. et al. (1998) The unified medical language system: an informatics research collaboration. *J. Am. Med. Inform. Assoc.*, 5, 1–11.
- Piñero, J. et al. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, 48, D845–D855.
- Raybould, M. et al. (2021) CoV-AbDab: the coronavirus antibody database. *Bioinformatics*, 37, 734–735.
- Reese, J. et al. (2021) KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns*, 2, 100155.
- Ren, J. et al. (2018) iTextMine: integrated text-mining system for large-scale knowledge extraction from the literature. *Database*, 2018, btaa834.
- Roseblat, G. et al. (2013) A methodology for extending domain coverage in SemRep. *J. Biomed. Inf.*, 46, 1099–1107.
- Szklarczyk, D. et al. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, 47, D607–D613.
- UniProt Consortium. (2020) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, 49, D480–D489.
- Wang, L. et al. (2020) COVID-19: the Covid-19 open research dataset. In: *ACL NLP-COVID Workshop 2020*, July 09–10, Seattle, Washington, USA.
- Wei, C. et al. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, 47, W587–W593.
- Wilkinson, M. et al. (2016) The fair guiding principles for scientific data management and stewardship. *Sci. Data*, 3, 160018.
- Wishart, D. et al. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46, D1074–D1083.
- Zhang, R. et al. (2021) Drug repurposing for COVID-19 via knowledge graph completion. *J. Biomed. Inf.*, 115, 103696.
- Zhu, F. et al. (2021) 2021 update on the clinical management and diagnosis of Kawasaki disease. *Curr. Infect. Dis. Rep.*, 23, 3.