*Bioinformatics*, 38(20), 2022, 4843–4845
https://doi.org/10.1093/bioinformatics/btac592
Advance Access Publication Date: 30 August 2022
Applications Note

Databases and ontologies

# CovidGraph: a graph to fight COVID-19

**Lea Gütebier [1,2], Tim Bleimehl[2,3], Ron Henkel [1], Jamie Munro[2], Sebastian Müller[2], Axel Morgner[2], Jakob Laenge[2], Anke Pachauer[2], Alexander Erdl[2], Jens Weimar[2], Kirsten Walther Langendorf[2], Vincent Vialard[2], Thorsten Liebig [2], Martin Preusse[2], Dagmar Waltemath [1] and Alexander Jarasch [2,3,*]**

[1]Medical Informatics Laboratory, University Medicine Greifswald, Greifswald 17475, Germany, [2]HealthECCO, Freiburg 79098, Germany and [3]German Center for Diabetes Research (DZD), Neuherberg 85764, Germany

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 6, 2022; revised on May 20, 2022; editorial decision on August 19, 2022; accepted on August 29, 2022

## Abstract

**Summary:** Reliable and integrated data are prerequisites for effective research on the recent coronavirus disease 2019 (COVID-19) pandemic. The CovidGraph project integrates and connects heterogeneous COVID-19 data in a knowledge graph, referred to as 'CovidGraph'. It provides easy access to multiple data sources through a single point of entry and enables flexible data exploration.

**Availability and Implementation:** More information on CovidGraph is available from the project website: https://healthecco.org/covidgraph/. Source code and documentation are provided on GitHub: https://github.com/covidgraph.

**Contact:** contact@healthecco.org

**Supplementary information:** Supplementary data is available at *Bioinformatics* online.

## 1 Introduction

In 2019, a novel coronavirus emerged causing the worldwide coronavirus disease 2019 (COVID-19) pandemic. The SARS-CoV-2 virus, or severe acute respiratory syndrome coronavirus 2, infects humans and causes the life-threatening coronavirus disease. A lot of data, including biomedical research data, has become available since the outbreak. Collecting and connecting these data are essential in the understanding of the virus and the fight against the pandemic.

COVID-19-related data, e.g. genes, transcripts and proteins, quickly became available from different data domains. The data items are often described or used in publications, which themselves are accessible from different databases and repositories. In addition, patents, clinical trials and computational models predicting the spread of the disease have emerged. Consequently, researchers wishing to investigate COVID-19 causes, spread and related diseases need to know and consult multiple data domains. Due to heterogeneous data formats and a lack of common interfaces, this data collection process is time consuming and error prone. However, specifically during a pandemic, policymakers and health care providers need to act quickly to control the virus. Hence, a fast access, a correct integration and a central access point are needed.

The CovidGraph project implements these requirements and integrates COVID-19-related data into a knowledge graph, thus allowing for rapid and intuitive data exploration across different data domains. It is a project actively developed and maintained by HealthECCO (https://healthecco.org), an organization that focuses on data integration in the fields of medicine and biology. It was founded in 2020 and connects a team of professionals from research, data science and medicine who aim to help making data related to COVID-19 freely and easily accessible for researchers, health workers and policymakers in order to foster collaborative research.

## 2 CovidGraph

The CovidGraph project works towards creating a central hub of connected information serving as an overview and access point for important SARS-CoV-2 knowledge. As of November 2021, the knowledge graph contains publications, patents, clinical trials, biomedical data such as genes and other molecular data, statistical and geographical information, and systems biology data represented in a Neo4j graph database (https://neo4j.com/) with approximately 36 million nodes and 60 million relationships, continuously growing.

Information about *publications*, including authors and their affiliation, are integrated from the COVID-19 Open Research Dataset (CORD-19), a collection of COVID-19 and coronavirus-related research papers (Wang *et al.*, 2020). Information about *patents* related to COVID-19 is obtained from 'The Lens' (https://www.lens.org/). The public registry ClinicalTrials.gov (Zarin *et al.*, 2011) serves as the data domain for *clinical trials* that are related to COVID-19. *Biomedical entities* such as genes, transcripts and proteins are integrated from well-known genome databases, among others, Ensembl (Hubbard *et al.*, 2002), NCBI RefSeq (O'Leary *et al.*, 2016) and

UniProt (The UniProt Consortium, 2021). Also included is information about pathways and gene expression data. Moreover, CovidGraph integrates functional annotation data from relevant ontologies, such as the Disease Ontology (Cowell and Smith, 2010). In addition, data are provided by the Hetionet resource (Himmelstein *et al.*, 2017), an integrative network of biomedical data. *Statistical data* such as case numbers is integrated from Johns Hopkins University (Dong *et al.*, 2020). The United Nations World Population Prospects 2019 (https://population.un.org/wpp/) offers population estimates and projections which are also integrated into CovidGraph. *Systems biology data*, in this case, simulation models and associated meta-data are integrated from MaSyMoS (Henkel *et al.*, 2015), which is a knowledge graph itself. MaSyMoS contains graph representations of models and meta-data from BioModels (Malik-Sheriff *et al.*, 2020), including a collection containing reproducible simulation studies of COVID-19 models (https://www.ebi.ac.uk/biomodels/covid-19).

A knowledge graph is most usable if, indeed, the knowledge is connected. Therefore, mappings across data domains are specified and these connections implemented in CovidGraph. One example of connected domains is literature and ontological knowledge. The corpus of included *publications*, *patents* and *clinical trials* is processed and occurrences of biomedical terms are recognized. Subsequently, recognized terms are mapped to their corresponding biomedical ontology entries (Gütebier *et al.*, 2021).

## 3 Interfaces and availability

CovidGraph enables easy browsing across different data sources. Data exploration can be started via one of four user interfaces or programmatically via an API. Each interface is tailored towards a specific exploration approach. The interfaces as described below can be accessed through the CovidGraph website (https://healthecco.org/covidgraph/).

The *Visual Graph Explorer* by yWorks provides a variety of predefined views for an intuitive keyword-based graph exploration (Supplementary Fig. S2). No prior knowledge of database query languages or the underlying graph structure is necessary for this straightforward approach. Users can interact with the clear interface by entering keywords in the search bar and filtering them for entities such as papers, patents, genes and more. Results are displayed as easily understandable connected glyphs allowing for a customized view of the selected data. When selecting a glyph (e.g. the gene *ACE2*), the user can load all related information (e.g. all encoded proteins). On request, additional information about the returned glyphs is provided in the detail panel. Moreover, the user can retrieve the underlying database queries for additional exploration options.

*SemSpect* (Liebig *et al.*, 2017) provided by derive (https://www.derivo.de/) offers unfiltered access to the database via an interactive drag-and-drop exploration tool (Supplementary Fig. S1). Data items selected by the user are automatically aggregated in groups (e.g. synonyms for the gene *ACE2*), and relations between different groups are displayed in an expandable tree structure. The user can filter, select and highlight single nodes inside a group (e.g. the gene *ACE2*), thereby building a tailored visual representation of the data without detailed knowledge of the underlying graph structure or query language. The resulting, customized visualization can be exported.

Experienced users may access the database directly through Cypher queries and in combination with visual exploration, offering more sophisticated access to the data. The interface is provided by the build-in *Neo4j Browser*.

More precisely, the integrated command line facilitates pattern matching on the graph database by entering specific queries for pattern and paths. Query results are returned as a visual representation of the resulting graph or, depending on the user request, as a tabular format or attribute-value pairs. The interface also includes the Graph Data Science and Awesome Procedures On Cypher libraries (https://neo4j.com/developer/graph-platform/). Both allow the user to natively use common graph algorithms (e.g. Dijkstra) or a variety of data import, export and manipulation methods, respectively.

*Neo4j Bloom* is an easy-to-use graph exploration application for visually interacting with Neo4j databases (Fig. 1). Neo4j Bloom provides easy-to-configure graph visualization, as well as rule-based styling for nodes or relationships. Using semi-natural language

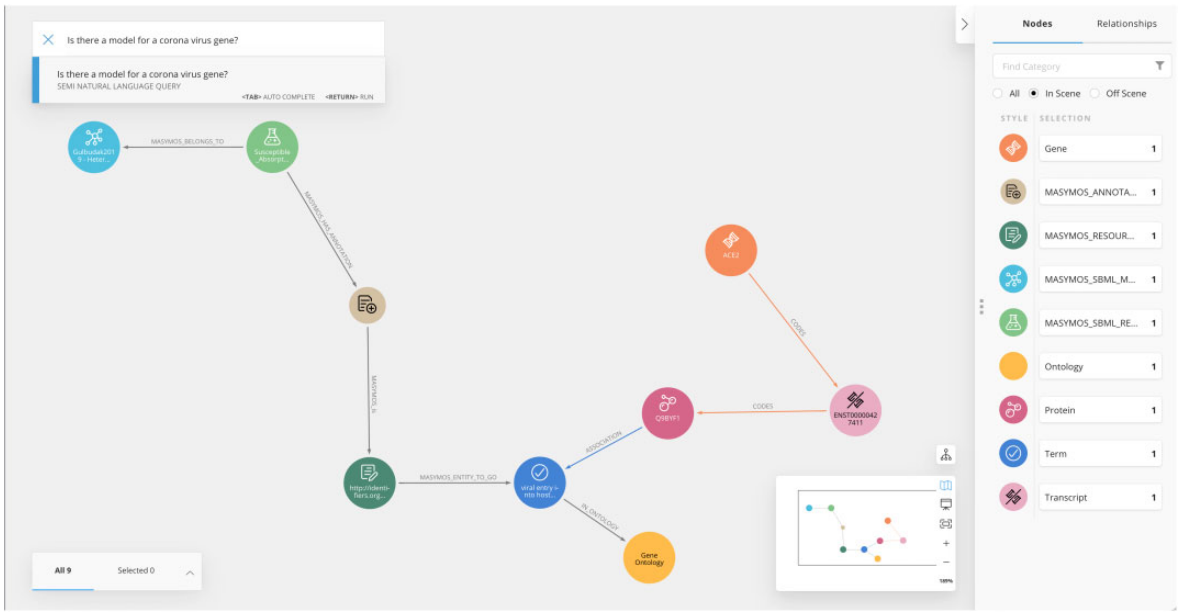


**Fig. 1.** Screenshot of Neo4j Bloom with semi-natural language query 'Is there a model for a corona virus gene?'. This figure shows the shortest path between the systems biology model 'Gulbudak2019.1—Heterogeneous viral strategies promote coexistence in virus-microbe systems (Lytic)' (cyan, BioModels, BIOMD0000000845) by Gulbudak and Weitz (2019) which is connected to MaSyMoS_Reaction (light green), MaSyMoS_Annotation (ochre), MaSyMoS_Resource (green) to the Gene Ontology term 'Virus entry into host cell' (blue, GO: 0019063) which, in turn, is associated with the protein Q9BYF1 (plum, UniProt) coded by the transcript ENST00000427411 (pink, RefSeq) of the corona virus gene *ACE2* (orange, GenBank, ENSG00000130234) (A color version of this figure appears in the online version of this article.)

queries (https://neo4j.com/product/bloom/) allows non-computer scientists researchers to easily query the graph databases by typing phrases and sentences in the search bar.

In addition, the Neo4j API is available for a programmatic access. It enables connection with external tools and programs for automatic querying or downloading of the data.

## 4 Use case

The cell surface receptor that is utilized by SARS-CoV-2 to enter the host cells is called angiotensin-converting enzyme 2 (*ACE2*). During the cell entry, the receptor acts as a binding site for the spike protein of the coronavirus Ni *et al.* (2020). To gain more knowledge and a better understanding about the mechanisms involved in a coronavirus infection, we want to investigate the underlying biological processes. CovidGraph connects important knowledge about the *ACE2* receptor with existing simulation models. Simulation models can assist in research by describing a biological system in a machine-readable format and enabling computational simulation and analysis of the system.

For our use case, we run the predefined query 'Are there systems biology models for the gene *ACE2*?' in Bloom which returns a set of fitting simulation models within seconds. Figure 1 shows one of the identified models and the connecting path between the data domains. Next, we identify suitable models within the resulting set to investigate the above-mentioned mechanisms underlying a coronavirus infection. Therefore, we again query the data in CovidGraph. The query 'Which models are related to coronavirus entry into host cell?' makes use of the ontological term 'viral entry into host cell' represented in a node and returns relevant models. The returned models include three models on HIV infection, a virus that enters the host cell via membrane fusion, and a model on influenza virus replication describing a receptor-mediated endocytosis (Heldt *et al.*, 2012). Membrane fusion and endocytosis are two main mechanisms for viral entry into cells (Jackson *et al.*, 2022). Based on previous knowledge, these models can help to gain insights into the entry mechanisms of the coronavirus.

The use of open data and open software tools, including easy access and exploration of the integrated, heterogeneous COVID-19 data, makes the CovidGraph a FAIR resource (Wilkinson *et al.*, 2016). The user is provided with a *F*indable, *A*ccessible, *I*nteroperable, *R*euseable resource that can be easily and intuitively explored with the aforementioned interfaces. CovidGraph's open-source policy and free accessibility allow for an unrestricted usage and, in fact, for integration of further data sources. Its modular framework encourages the addition of new data sources and supports periodical updates of the integrated data, an update strategy we are currently implementing.

## Acknowledgements

The work presented here is the work of the HealthECCO team.

## Funding

The COVID-19 collection on BioModels was developed as part of Grant: European Commission: EOSCsecretariat.eu—EOSCsecretariat.eu (831644).

*Conflict of Interest*: none declared.

## References

Cowell,L.G. and Smith,B. (2010). Infectious disease ontology. In: Sintchenko, V. (ed) *Infectious Disease Informatics*. Springer, New York, pp. 373–395. https://doi.org/10.1007/978-1-4419-1327-2_19.

Dong,E. *et al.* (2020) An interactive web-based dashboard to track COVID-19 in real time. *Lancet. Infect. Dis.*, **20**, 533–534.

Gulbudak,H. and Weitz,J.S. (2019) Heterogeneous viral strategies promote coexistence in virus-microbe systems. *J. Theor. Biol.*, **462**, 65–84.

Gütebier,L. *et al.* (2021) COVIDGraph: connecting biomedical COVID-19 resources and computational biology models. In: 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores, SEA-Data 2021. pp. 34–37.

Heldt,F.S. *et al.* (2012) Modeling the intracellular dynamics of influenza virus replication to understand the control of viral RNA synthesis. *J. Virol.*, **86**, 7806–7817.

Henkel,R. *et al.* (2015) Combining computational models, semantic annotations and simulation experiments in a graph database. *Database*, **2015**, 1–6.

Himmelstein,D.S. *et al.* (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, **6**, e26726.

Hubbard,T. *et al.* (2002) The ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Jackson,C.B. *et al.* (2022) Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.*, **23**, 3–20.

Liebig,T. *et al.* (2017). Connecting the Dots in Million-Nodes Knowledge Graphs with SemSpect. In: *International Semantic Web Conference (Posters, Demos & Industry Tracks)*.

Malik-Sheriff,R.S. *et al.* (2020) BioModels - 15 years of sharing computational models in life science. *Nucleic Acids Res.*, **48**, D407–D415.

Ni,W. *et al.* (2020) Role of angiotensin-converting enzyme 2 (ACE2) in COVID-19. *Crit. Care*, **24**, 1–10.

O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.

Wang,L.L. *et al.* (2020). CORD-19: the Covid-19 Open Research Dataset. *arXiv*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/.

Wilkinson,M.D. *et al.* (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, **3**, 1–9.

Zarin,D.A. *et al.* (2011) The ClinicalTrials.gov results database - update and key issues. *N Engl. J. Med.*, **364**, 852–860.