Bioinformatics, 37(9), 2021, 1332–1334
doi: 10.1093/bioinformatics/btaa834
Advance Access Publication Date: 6 December 2020
Applications Note

Systems biology

COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology

Daniel Domingo-Fernández [1,2,*], Shounak Baksi[3,*], Bruce Schultz [1,*], Yojana Gadiya [1,2], Reagon Karki[1,2], Tamara Raschka[1,2], Christian Ebeling[1], Martin Hofmann-Apitius [1,2] and Alpha Tom Kodamullil [1,2,*]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany, [2]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany and [3]Causality Biomodels, KINFRA Hi-Tech Park, Cochin, Kerala 683503, India

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on April 14, 2020; revised on July 26, 2020; editorial decision on September 9, 2020; accepted on September 11, 2020

Abstract

Summary: The COVID-19 crisis has elicited a global response by the scientific community that has led to a burst of publications on the pathophysiology of the virus. However, without coordinated efforts to organize this knowledge, it can remain hidden away from individual research groups. By extracting and formalizing this knowledge in a structured and computable form, as in the form of a knowledge graph, researchers can readily reason and analyze this information on a much larger scale. Here, we present the COVID-19 Knowledge Graph, an expansive cause-and-effect network constructed from scientific literature on the new coronavirus that aims to provide a comprehensive view of its pathophysiology. To make this resource available to the research community and facilitate its exploration and analysis, we also implemented a web application and released the KG in multiple standard formats.

Availability and implementation: The COVID-19 Knowledge Graph is publicly available under CC-0 license at https://github.com/covid19kg and https://bikmi.covid19-knowledgespace.de.

Contact: daniel.domingo.fernandez@scai.fraunhofer.de or shounak.baksi@causalitybiomodels.com or bruce.schultz@scai.fraunhofer.de

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

1 Introduction

The COVID-19 crisis has prompted a response of the scientific community that is unparalleled in history. Research organizations have dedicated their entire workforce to combat the pandemic. Tens of thousands of researchers in hundreds of universities, governmental laboratories and industrial research departments have entirely focused their efforts on understanding the virus pathophysiology, finding drugs that interfere with its life cycle and developing immunization strategies for future vaccines (Chahrour *et al.*, 2020).

While the steep increase in research activities in the COVID-19 context has led to an unprecedented increase of scientific publications, it becomes challenging to identify genuine novel findings and discern them from those that are already known. The process of discriminating ‘known knowns’ from ‘unknown knowns’ can be supported by knowledge graphs (KGs), as they provide a means to capture, represent and formalize structured information (Nelson *et al.*, 2019). Furthermore, although these KGs were originally developed to describe interactions between entities, they are complemented by a broad range of algorithms that have been proven to partially automate the process of knowledge discovery (Cowen *et al.*, 2017; Humayun *et al.*, 2020). Importantly, novel machine learning techniques can generate latent, low-dimensional representations of the KG which can then be utilized for downstream tasks such as clustering or classification (Hamilton *et al.*, 2017).

In this article, we present an approach to lay the foundation for a comprehensive KG in the context of COVID-19. Our work is complemented by a web application that enables users to comprehensively explore the information contained in the KG. To facilitate the ease of usage and interoperability of our KG, we have released its content in various standard formats to promote its adoption and enhancement by the scientific community.

2 Material and methods

In this section, we outline the methodology used to: (i) select the corpus, (ii) generate the COVID-19 KG and (iii) develop the web application for exploring the KG.

2.1 Selection of scientific literature

For the creation of the KG, scientific literature related to COVID-19 was retrieved from open access and freely available journals (see details in Supplementary Text). This corpus was then filtered based on available information about potential drug targets for COVID-19, biological pathways in which the virus interferes to replicate in its human host, and information on the various viral proteins along with their functions. Finally, the articles were prioritized based on the level of information that could be captured in the modeling language used to build the KG.

2.2 Constructing the COVID-19 Knowledge Graph

Evidence text from the prioritized corpus was manually encoded in Biological Expression Language (BEL) as a triple (i.e. source node—relation—target node) including metadata about the nodes and their relationships as well as corresponding provenance and contextual information. BEL scripts generated from this curation work are freely available at https://github.com/covid19kg along with their network representations in several other standard formats (e.g. SIF, GraphML and NDEx). By making this data available in multiple formats, we are seeking to facilitate the analysis of the KG with a broad range of methods/software as well as promote its integration into other biological databases and web services such as the one presented in the following section.

2.3 Web application

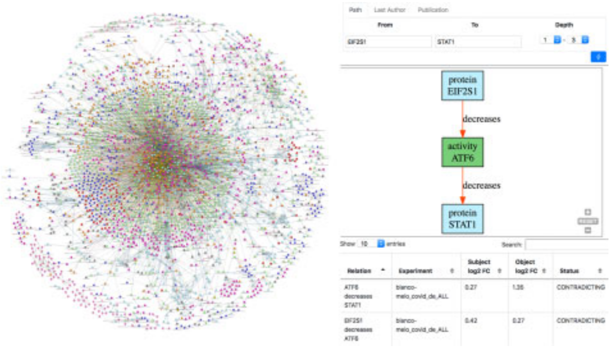
To better aid the exploration and usage of the generated COVID-19 Knowledge Graph, a web application was developed using Biological Knowledge Miner (BiKMi), an in-house software package designed for exploring pathways and molecular interactions within a BEL-derived network. The front-end of the application was constructed using the Python Django web framework, while the back-end of the software is implemented using OrientDB, a multi-model database management system that allows for both relational and graph queries to be made against a database via its API (Supplementary Text), which opens the avenue towards systematic comparison of different COVID models.

3 Results

We introduce a KG that comprises mechanistic information on COVID-19 published in 160 original research articles. In its current state, the COVID-19 KG incorporates 4016 nodes, covering 10 entity types (e.g. proteins, genes, chemicals and biological processes) and 10 232 relationships (e.g. increases, decreases and association), forming a seamless interaction network (Supplementary Text). Given the selected corpora, these cause-and-effect relations primarily denote host-pathogen interactions as well as comorbidities and symptoms associated with COVID-19. Furthermore, the KG contains molecular interactions related to host invasion (e.g. spike glycoprotein and its interaction with the host via receptor ACE2) and the effects of the downstream inflammatory, cell survival and apoptosis signaling pathways.

A key aspect of the COVID-19 KG is in its large coverage of drug–target interactions along with the biological processes, genes and proteins associated with the novel coronavirus. We have identified over 300 candidate drugs currently being investigated in the context of COVID-19 (Supplementary Text), including proposed repurposing candidates and drugs under clinical trial.

Along with the KG, we implemented a web application (https://bikmi.covid19-knowledgespace.de) for querying, browsing and navigating the KG (Fig. 1). The visualization enables users to explore and query the network (e.g. filtering nodes or edges, or calculating paths between nodes of interest). Additionally, it enables users to upload *omics* data and validate its signals against the knowledge contained in the network. To demonstrate this feature, the web application is loaded with the transcriptomics experiments conducted by Blanco-Melo *et al.* (2020).



From: EIF2S1 To: STAT1 Depth: 1 - 3

Relation	Experiment	Subject log2 FC	Object log2 FC	Status
ATF6 decreases STAT1	blanco-melo_covid_de_ALL	0.27	1.38	CONTRADICTING
EIF2S1 decreases ATF6	blanco-melo_covid_de_ALL	0.42	0.27	CONTRADICTING

Fig. 1. (Left) Visualization of the COVID-19 KG in BiKMi. (Right) Querying paths between two nodes and verifying their consistency with transcriptomics data

4 Discussion

The novel coronavirus has motivated a profound response by the scientific community and has led to the rapid publishing of COVID-19 research. As an attempt to organize and formally represent the most current knowledge of the virus, we have introduced a KG comprising mechanistic information around COVID-19 biology and pathophysiology. The presented KG provides a comprehensive overview of relevant viral protein interactions and their downstream molecular mechanisms. Additionally, it also includes the vast majority of potential drug–targets as well as clinical manifestations associated with comorbidities and symptoms. Given the biological complexity and the sparse information we currently have on the pathophysiology of the virus, mechanistic knowledge contained in the KG could be promising for the discovery of yet hidden interactions. The COVID-19 KG presented here is part of a bigger ecosystem that integrates disease maps with three of the largest pathway databases (Domingo-Fernández *et al.*, 2019).

Not only do we provide a web application to make the content accessible to the research community, but we also have released the KG in a variety of standard formats. In doing so, we aim to foster an exchange of information across similar modeling approaches (Ostaszewski *et al.*, 2020) (Supplementary Text) as well as to facilitate its analytic use on both knowledge- and data-driven methods. Furthermore, the knowledge present in high-quality manually curated approaches can be combined with the information extracted by scalable text mining approaches such as Wang *et al.* (2020), which enable to systematically scan COVID-19 literature and construct KGs based on entity co-occurrence or relation extraction. However, combining both modeling approaches involves understanding their differences as well as relative strengths and weaknesses, some of which are discussed in Supplementary Text. Finally, we plan to make future releases of the KG to ensure the most up-to-date content as well as to benefit from its integration and crosstalk with other similar activities (i.e. #covidpathways).

Funding

This work was supported by the MAVO and ICON programs of the Fraunhofer Society.

Conflict of Interest: none declared.

References

Blanco-Melo,D. *et al.* (2020) Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, **181**, 1036–1045.

Chahrour,M. *et al.* (2020) A bibliometric analysis of COVID-19 research activity: a call for increased output. *Cureus*, **12**, e7357.

Cowen,L. *et al.* (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.*, **18**, 551–562.

Domingo-Fernández,D. *et al.* (2019) PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, **20**, 243.

Hamilton,W.L. *et al.* (2017) Representation learning on graphs: methods and applications. *IEEE Data Eng. Bull.*, **40**, 52–74

Humayun,F. *et al.* (2020) A computational approach for mapping heme biology in the context of hemolytic disorders. *Front. Bioeng. Biotechnol.*, **8**, 74.

Nelson,W. *et al.* (2019) To embed or not: network embedding as a paradigm in computational biology. *Front. Genet.*, **10**, 381.

Ostaszewski,M. *et al.* (2020) COVID-19 disease map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci. Data*, **7**, 1–4.

Wang,Q. *et al.* (2020) COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. https://www.cell.com/cell/fulltext/S0092-8674(20)30489-X?_return.