COVID-KOP: Integrating Emerging COVID-19 Data with the ROBOKOP Database

Daniel Korn[1], Tesia Bobrowski[2], Michael Li[1], Yaphet Kebede[3], Patrick Wang[4], Phillips Owen[3], Gaurav Vaidya[3], Eugene Muratov[2], Rada Chirkova[5], Chris Bizon[3]*, and Alexander Tropsha[2*].

[1]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7568, USA; [2]School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7568, USA; [3]Renaissance Computing Institute, University of North Carolina at Chapel Hill, NC 27599-7568, USA; [4]CoVar Applied Technologies, Durham, NC 27701, USA; and [5]Department of Computer Science, North Carolina State University, Raleigh, NC, 27606-5550.

Contact: bizon@renci.org and alex_tropsha@unc.edu

`

Abstract

In response to the COVID-19 pandemic, we established COVID-KOP, a new knowledgebase integrating the existing ROBOKOP biomedical knowledge graph with information from recent biomedical literature on COVID-19 annotated in the CORD-19 collection. COVID-KOP can be used effectively to test new hypotheses concerning repurposing of known drugs and clinical drug candidates against COVID-19. COVID-KOP is freely accessible at https://covidkop.renci.org/. For code and instructions for the original ROBOKOP, see: https://github.com/NCATS-Gamma/robokop.

Introduction

In the absence of effective medications for COVID-19, there is an urgent need to identify drugs that can combat this ongoing pandemic. This task can be accomplished most rapidly by repurposing the existing medications. Biomedical knowledge graphs such as ROBOKOP (Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways) developed by our team recently [1] provide an efficient way to identify potential candidate drugs by making inferences upon the relationships between knowledge graph nodes. We have merged the ROBOKOP knowledge graph with the new supply of COVID-19 related information from recent publications and other knowledge sources to form COVID-KOP.

Methods

We employed COVID-19 Open Research Dataset (CORD-19, https://allenai.org/data/cord-19) containing over 60,000 full-text research papers that can be used and redistributed for studies on COVID-19. An ontological tagging of the CORD-19 dataset was provided in GitHub for public use by the SciBiteAI group (https://github.com/SciBiteLabs/CORD19). We parsed CORD-19 data into a format compatible with the ROBOKOP's knowledge graph by extracting, sentence by sentence, the counts of ontological terms and tags co-occurrences. This created 800,000 new edges in the COVID-KOP knowledge graph. Additionally, we employed the SciGraph tool (https://github.com/SciGraph/SciGraph), which also allows biomedical ontological term tagging and tag co-occurrence counts at the paper rather than sentence level, leading to 4.5 million new edges.

Gene Ontology Annotation data for all viral proteins, including those of SARS-CoV-2, were downloaded from the EBI FTP site (see https://github.com/TranslatorIIPrototypes/ViralProteome for details). The knowledge graph integration tool KGX (https://github.com/NCATS-Tangerine/kgx) was used to merge the GOA data and create a ROBOKOP-formatted graph. In total, the COVID-KOP database and knowledge graph comprise nodes for 40,000 proteins, 4,000 NCBITaxon [2] terms, 1,300 GO annotations [3], and 232,000 new edges on top of those in ROBOKOP.

A set of 26 SARS-CoV-2 symptoms was identified from various resources (https://www.cebm.net/covid-19/covid-19-signs-and-symptoms-tracker/; https://covid.cd2h.org/N3C; https://www.hematology.org/covid-19/covid-19-and-coagulopathy) and a recent commentary [4]. This information was not in a convenient format for scraping (images, small tables, etc.) and so it was manually entered into the COVID-KOP database as edges between the COVID-19 and the phenotypes.

Due to multiple identifier systems that different databases used to refer to the same entities, we utilized the Data Translator Node Normalization API (https://github.com/TranslatorIIPrototypes/NodeNormalization) for data integration. COVID-KOP is powered by the knowledge graph database Neo4J (https://neo4j.com/), which employs Cypher language to enable complex graph database queries.

New data were added to COVID-KOP using Python scripts, which iterated through all novel entries and added them to the ROBOKOP KG as nodes. Then all newly discovered connections were added to the KG as new edges and given the label "related_to". This fully integrated COVID-KOP KG can be mined in the same way as ROBOKOP KG [1].

Case Study

We illustrate the utility of COVID-KOP by examining the linagliptin – COVID-19 connection (Figure 1). Linagliptin is a drug for type 2 diabetes (T2D) that is currently undergoing clinical trials for COVID-19 (https://clinicaltrials.gov/ct2/show/NCT04341935). Linagliptin inhibits dipeptidyl peptidase-4 (DPP-4), which degrades hormones stimulating insulin production [5]. Moreover, DPP-4 is known to be overexpressed in patients with T2D [6]. COVID-19 patients with T2D have a higher risk of developing more severe symptoms, possibly due to an increased expression of the host receptor ACE2 [7].

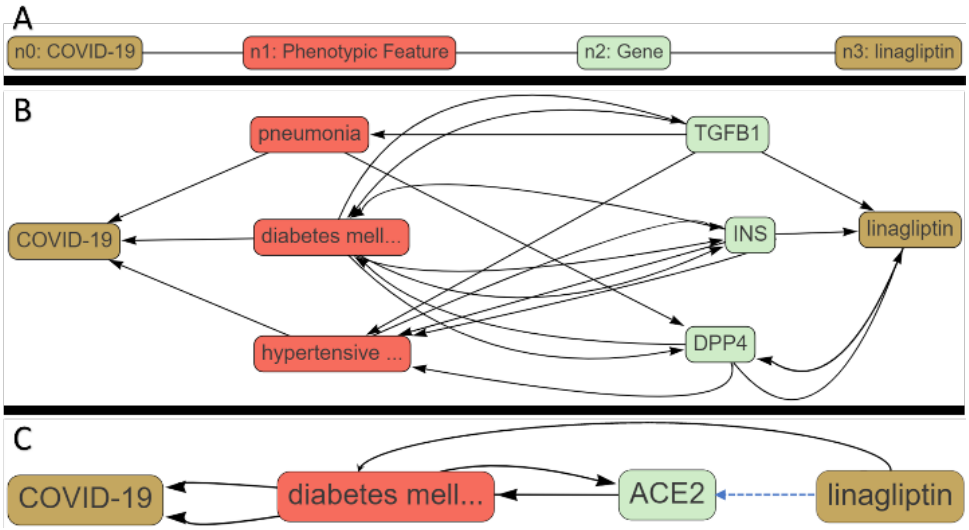

Figure 1. *Execution of a COVID-KOP query for linagliptin-COVID-19 pair. (A) query graph; (B) answer graph for linagliptin-DPP4-T2D-COVID-19 pathway; (C) answer graph for linagliptin-ACE2-T2D-COVID-19 pathway.*

We applied COVID-KOP to identify possible mechanistic connections between linagliptin and COVID-19 that could either support its expected therapeutic effect or identify possible complications this drug may cause in COVID-19 patients. A query was constructed from individual biomedical objects using the COVID-KOP user interface. Individual nodes are placed and linked to specific biomedical objects (such as node **n2** being a **gene** in Figure 1a). The user may then choose to link these nodes together in any order they choose. In this case study, node **n0** is matched to COVID-19; **n1** is marked as any phenotypic feature linked to COVID-19; and **n2** is any gene related to a phenotypic feature connected to COVID-19. Finally, **n2** must also have a connection to the linagliptin.

By running this query, we quickly retrieved the pathway serving as a rationale for the linagliptin clinical trial against COVID-19 (Linagliptin-T2D-DPP4-COVID-19; Figure 1B). Three conditions – pneumonia, diabetes mellitus II, and hypertensive disorder – were identified as associated with COVID-19. Genes associated with these conditions that are also related to linagliptin are annotated in Figure 1B as well. This answer graph suggests that the use of linagliptin may inhibit TGFβ-1 transcription and possibly increase patients' risk of developing more severe pneumonia, even though it may alleviate some of the more severe pathologies of COVID-19 seen in T2D patients due to inhibition of DPP-4.

Using COVID-KOP, we also uncovered an additional inference for associating linagliptin and Angiotensin-Converting Enzyme II (ACE2), a host receptor that SARS-CoV-2 uses to enter cells: Linagliptin-ACE2-T2D-COVID-19 (Figure 1C). Expression of ACE2 is increased in patients with T2D; thus, ACE inhibitors and angiotensin-receptor blockers are commonly used to treat individuals with this condition [8]. It is yet unclear if upregulating ACE2 would be beneficial or detrimental to patients [8,9], especially those with T2D; nevertheless, this inference reveals another possible pathway by which linagliptin could influence COVID-19 patient outcomes. This case study illustrates how COVID-KOP could help recover both known biochemical pathways associating a drug with COVID-19 (linagliptin-DPP4-Diabetes-Type2-COVID-19) as well as offer potentially novel inferences (linagliptin-TGFB1-Pneumonia-COVID-19).

Conclusions

In response to the COVID-19 epidemic, we developed COVID-KOP, a knowledgebase and a web portal that integrates the existing ROBOKOP biomedical knowledge graph with information gathered from recently published biomedical information regarding COVID-19. The case study described here illustrates the utility of COVID-KOP in uncovering both known and unknown inferences between the drugs and COVID-19, which can lead to the development or preliminary screening of new or existing chemotherapies for COVID-1.

Funding

This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health [OT2R002514] and National Institutes of Health [1U01CA207160].

Conflict of Interest: none declared.

References

(1) Bizon, C.; Cox, S.; Balhoff, J.; Kebede, Y.; Wang, P.; Morton, K.; Fecho, K.; Tropsha, A. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *J. Chem. Inf. Model.* **2019**, *59*, 4968–4973.

(2) Federhen, S. The NCBI Taxonomy Database. *Nucleic Acids Res.* **2012**, *40*, D136–D143.

(3) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25*, 25–29.

(4) Schett, G.; Sticherling, M.; Neurath, M. F. COVID-19: Risk for Cytokine Targeting in Chronic Inflammatory Diseases? *Nature Reviews Immunology*. Nature Research May 2020, pp 271–272.

(5) Scott, L. J. Linagliptin: In Type 2 Diabetes Mellitus. *Drugs* **2011**, *71*, 611–624.

(6) Barchetta, I.; Cavallo, M. G.; Baroni, M. G. COVID-19 and Diabetes: Is This Association Driven by the DPP4 Receptor? Potential Clinical and Therapeutic Implications. *Diabetes Res.*

Clin. Pract. **2020**, *163*, 108165.

(7) Fang, L.; Karakiulakis, G.; Roth, M. Are Patients with Hypertension and Diabetes Mellitus at Increased Risk for COVID-19 Infection? *Lancet Respir. Med.* **2020**, *8*, e21.

(8) Pal, R.; Bhansali, A. COVID-19, Diabetes Mellitus and ACE2: The Conundrum. *Diabetes Res. Clin. Pract.* **2020**, *162*, 108132.

(9) Nakhleh, A.; Shehadeh, N. Interactions between Antihyperglycemic Drugs and the Renin-Angiotensin System: Putative Roles in COVID-19. A Mini-Review. *Diabetes Metab. Syndr.* **2020**, *14*, 509–512.