A COVID-19 Knowledge Graph for Therapeutic Discovery from Semantic Integration of Literature and Databases

Karen E. Ross[1], Chuming Chen[2], Julie Cowart[2], Sachin Gavali[2], and Cathy H. Wu[1,2]

[1]Georgetown University Medical Center, Washington, DC; [2]University of Delaware, Newark, DE

Introduction: In response to the COVID-19 global health emergency, the COVID-19 literature is rapidly expanding. Computational approaches that automatically distill key information from text and integrate it with information from curated biological databases are essential to gain insight into COVID-19 etiology, diagnosis and treatment. Knowledge graphs (KGs) are a powerful way to represent such diverse biological information and generate novel hypotheses. In this study, we constructed a COVID-19 knowledge graph based on mining of literature and databases, using semantic web technologies (RDF and SPARQL) for data integration.

Methods: The KG integrates information extracted from (i) the COVID-19 literature using the text-mining tools iTextMine (PTM and miRNA relations), PubTator (biomedical entities), and SemRep (biomedical relations based on UMLS); (ii) curated databases, such as UniProtKB and DrugBank; and (iii) proteomic and phospho-proteomic data on SARS-CoV-2-infected cells. It is served by the OpenLink Virtuoso server community edition with SPARQL 1.1 query federation.

Results: The COVID-19 KG, consisting of 22 named graphs and 1.2 billion RDF triples, is accessible via a knowledge portal (https://research.bioinformatics.udel.edu/covid19kg/) with browsing and search interfaces; YASGUI (Yet Another Sparql GUI) with a set of comprehensive SPARQL queries for new users; and a RESTful API. Using the KG, we identified several potentially beneficial COVID-19 therapeutics, including drugs targeting TNF and IFN-gamma, two proteins implicated in the cytokine storm that affects some patients with severe COVID-19, as well as kinase inhibitors and miRNAs that may disrupt key molecular interactions of the SARS coronavirus nucleocapsid protein, a heavily phosphorylated protein required for viral genome replication and packaging.

Conclusions: With its unique focus on molecular relations, ability to keep up with the latest published results via text mining, and inclusion of a wide variety of biomedical knowledge using a semantic framework, our KG can provide insight into the rapidly evolving landscape of COVID-19.