Biomedical Knowledge Graph of COVID-19: Construction and Applications

Shuangjia Zheng[1,2], Jixian Zhang[1], Jiahua Rao[2], Xianglu Xiao[1], Wade Menpes-Smith[1], Jake Y. Chen[3], Yuedong Yang[2]* and Zhangming Niu[1]*

[1]Aladdin Healthcare Technologies Ltd., London EC1Y OUR, UK
[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510000, China
[3]Informatics Institute, the University of Alabama School of Medicine, Birmingham, AL
*Contact: zhangming@aladdinid.com; yangyd25@mail.sysu.edu.cn

ABSTRACT

As COVID-19 continues to spread rapidly across the world, it is urgent to find effective therapies to treat the pandemic. The White House and an alliance of leading research institutes have compiled the COVID-19 Open Research Dataset (CORD-19), containing over 51,000 research articles related to COVID-19, SARS-CoV-2, and coronaviruses. Although a large amount of data is made available, it remains challenging for researchers to go through each paper and identify critical information from vast bodies of unstructured texts. Automatic knowledge graph construction is an effective way to rapidly convert the complicated knowledge of COVID-19 into a structured graph format, which can then be utilized to downstream applications such as drug repositioning and mechanism analysis.

In this paper, we constructed a COVID-19 biomedical knowledge graph performed with minimum supervision on a flexible pipeline, which includes a series of steps, i.e., entity recognition, relation extraction, knowledge graph embedding, and its downstream application in drug repositioning (Figure 1). In particular, we first used a BERT-based model with semi-supervised biomedical facts as a seed set of knowledge to predict the relationship of each pair of entities that we collected with a template-based name entity recognition (NER) method from CORD-19. The constructed framework can effectively extract 39,583 structured facts (triplets) of 18 types relation with high precision, resulting in a COVID-19 specific knowledge graph with 47,031 relationships among genes, chemicals, and diseases (Figure 2). We then developed a neural attentive graph embedding model to map entities and relations into low-dimensional feature vectors while capturing their semantic meanings. We applied the embedding model to produce novel drug repositioning hypotheses and assessed their scientific validity using external sources. Finally, for high-scoring predictions, we analyzed the actual relation paths of the entity pairs and provided putative mechanistic interpretations (Figure 3).

KEYWORDS

COVID-19; Deep learning; Text mining; Knowledge graph; Drug repositioning.

BIOKDD'20, August 24, 2020, San Diego, CA

https://doi.org/

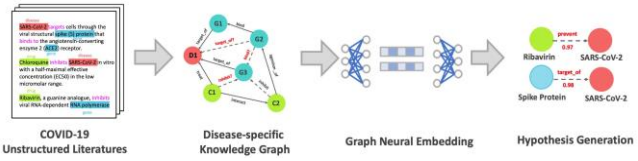

Figure 1. Schematic illustration of our automatic knowledge graph construction pipeline for discovering the potential drugs to treat the COVID-19 disease.

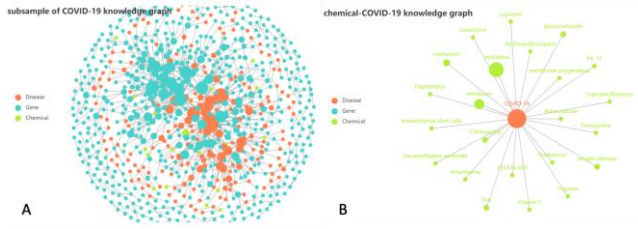

Figure 2. (A) Subsample of the COVID-19 knowledge graph. (B) Subsample of the chemical-COIVD-19 triplets.

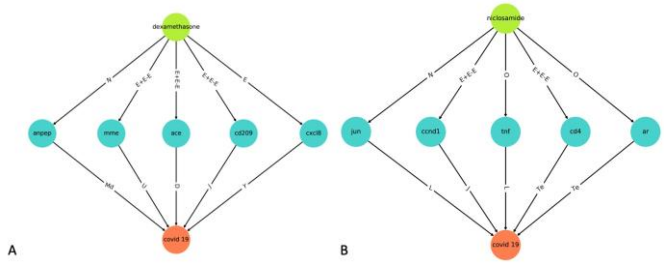

Figure 3. Mechanistic interpretations of two high-scoring triplets (Chemical, "Treat", "COIVD-19"). The treatment potential of (A) dexamethasone targeting COIVD-19 was scored as 0.755, while the (B) niclosamide was 0.635. Both of them were validated in recently published literatures (outside the training set).

Acknowledgments: This study was partially assisted via the financial grant from the Innovative Medicines Initiative Program – IMI2-RIA (proposal No: 101005122), the National Key R&D Program of China (2018YFC0910500), National Natural Science Foundation of China (U1611261, 61772566, and 81801132), and NIH Grant UL1 TR001417.