Phenotype-driven COVID-19 Knowledge Graph from Biomedical Literature Facilitates Hypothesis Generation

Ryan Lee (CAS ‘23, W ’23) , David Xu (CAS ’23), Jackie Peng, Yunyun Zhou, Kai Wang

Wang Genomics Lab, Children’s Hospital of Philadelphia, Philadelphia, PA 19104

Introduction

The Problem

- Starting December 2019, the COVID-19 pandemic has triggered a research ***“literature explosion”*** (Chahrour et al., 2020)
- Academics are struggling to ***differentiate*** knowns from unknowns and ***connect*** existing findings

Our Solution

- ***Knowledge graphs*** (KG) map unstructured information onto a network of nodes and edges (Figure A)
- KGs ***support multiple downstream tasks***, such as embedding, link prediction, and graph completion, that facilitate novel knowledge syntheses
- Past researchers have utilized KGs to display existing COVID-19 information but not for ***knowledge discovery***
- Our approach employs natural-language processing (NLP) and machine learning to construct, embed, and leverage a KG to ***predict potential COVID-19 treatments, risk factors, and symptoms***

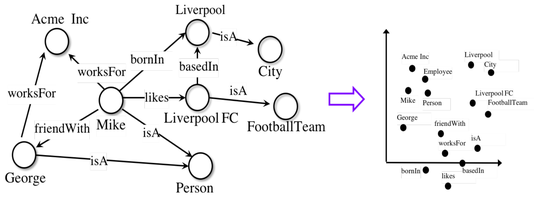


Figure A (above): An example knowledge graph depicts ideas as circular nodes and relationships as connecting lines. Algorithmically, components of this network can be mapped into a vector space, a process known as embedding (Accenture, 2020).

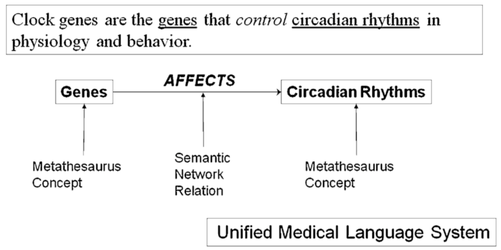
Materials & Methods

Step 1: Data Sourcing

- Sourced data from the ***COVID-19 Open Research Dataset*** (CORD-19) v22, a collection of 200,000+ scholarly texts related to COVID-19

Step 2: Triple Extraction

- ***SemRep:*** program by the National Institutes of Health that ***extracts triples from biomedical texts*** (subject-relation-object); see example (Rindflesch et al., 2011)



- Applied SemRep to corpus generated from CORD-19 to isolate triples involving ***genes, drugs, and symptoms***
- Included more entities and relations relevant to COVID-19 from an earlier list by Rashed et al. (2020)

Step 3: Knowledge Graph Creation & Embedding

- Imported triples into ***Neo4J***, a graph database system, along with paper authorship information to create and visualize the KG
- *KG Embedding:*
  - With the Trans-D approach from the OpenKE-Tensorflow1.0 package, ***triples embedded into a continuous vector space***
  - Evaluated embedding performance by feeding model “corrupted” triples with missing components and scoring its triple-completion predictions
  - Further trained model to process incomplete triples and ***rank predictions*** in descending order of likelihood and relevance

Results

Treatment Predictions	*Symptom Predictions*	*Risk Factor Predictions*
Chloroquine	**COVID-19**	**COVID-19**
Pharmaceutical Preparations	Cessation of Life	Diabetes
Antiviral Agents	**SARS**	Diabetes Mellitus
Iopinavir/Ritonavir	Complication	Cardiovascular Disease
Ribavirin	Disease	Chronic Disease
Hydroxychloroquine	Acute Kidney Failure	**SARS**
Azithromycin	Inflammatory Response	Angiotensin-Converting Enzyme Inhibitors
Nucleoside Analogs	Pneumonia	**Influenza**
Lead compound	Inflammation	Opportunistic Infections
Investigational new drugs	**Infection**	**CD69 Protein, Human**
Iopinavir	Vertical Disease Transmission	**ACE2 Gene**
Hydroxychloroquine Sulfate	Disorder of circulatory system	**Disease**

* bolded terms do not already exist in the KG

Figure B (left): At BFS depth of 1, the embedding model effectively predicts treatments, symptoms, and risk factors associated with COVID-19. Many of these new predictions are confirmed by promising literature, validating our methodology.

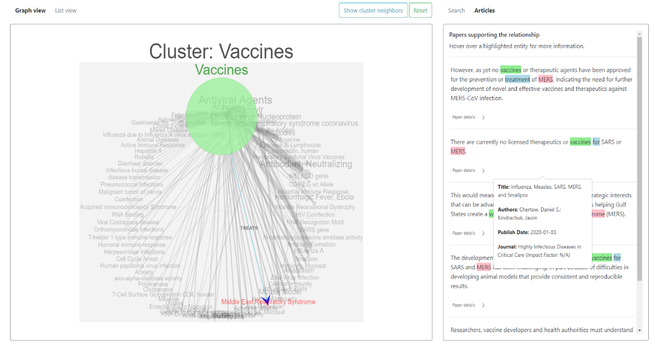


Figure C (above): Searching “vaccine” on our online interface returns the relevant cluster of nodes, paper excerpts, and authorship data.

Discussion

Conclusions

- Promising results reinforce KGs as effective tools to ***display*** and ***explore*** knowledge, especially for COVID-19
- In the context of COVID-19, our workflow uniquely provides a ***link-prediction mechanism*** and an online, ***user-friendly interface***

Future Directions

- Weight triples during KG creation based on their frequency to enhance and fine-tune link prediction
- Incorporate clinical phenotype notes to adapt KG for diagnostic purposes

Acknowledgements

Special Thanks To:

- Organizers, partners, and funders of the Penn Undergraduate Research Mentoring (PURM) Program
- Dr. Kai Wang, Dr. Yunyun Zhou, Jackie Peng, Siwei Xu, and other members of the Wang Genomics Laboratory of CHOP

Selected References

1. Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A. A., Salhab, H., Fares, M., & Khachfe, H. H. (2020). A Bibliometric Analysis of COVID-19 Research Activity: A Call for Increased Output. *Cureus, 12*(3). DOI: 10.7759/cureus.7357.
2. Accenture. (2020). *AmpliGraph*. AmpliGraph. https://docs.ampligraph/. org/en/1.2.0/index.html.
3. Rindflesch, Thomas & Kilicoglu, Halil & Fiszman, Marcelo & Rosemblat, Graciela & Shin, Dongwook. (2011). Semantic MEDLINE: An advanced information management application for biomedicine. Information Services and Use. 31. 15-21. 10.3233/ISU-2011-0627.