Poster 2205

Using Text Mining Link Prediction to Expedite COVID-19 Research

Soham Kulkarni[1], Kevin McCoy[1], Sai Sateesh Gudapati[1], Vivek Vanga[1], Jayant Prakash[1], Cassie S. Mitchell[1]

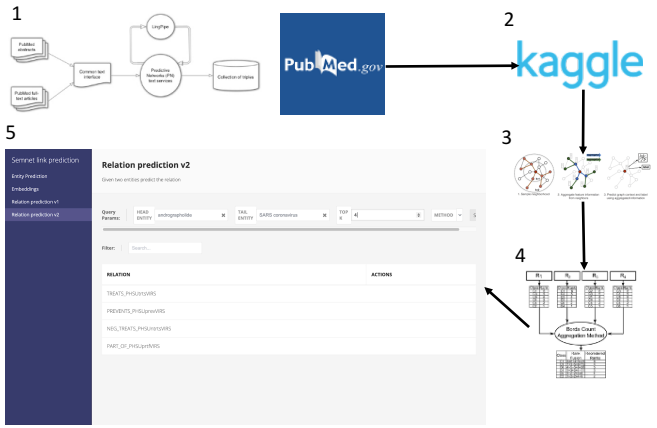
[1]Biomedical Engineering, Georgia Institute of Technology & Emory University, Atlanta, GA

Introduction

- SARS-CoV-2 is a novel disease that remains largely uncharacterized. FDA approved drugs have been repurposed with the goal of treating the symptoms of COVID-19 in order to save lives.
- Literature mining tools are able to generate "knowledge graphs" of relationships contained in the 30+ million abstracts in PubMed. Unfortunately, these literature mining tools only make predictions on known relationships.
- We have developed a new "link prediction" algorithm, which predicts new relationships connecting novel drug treatments and COVID-19. This link prediction subsequently ranks these novel relationships to COVID-19.

Methods

1. 30 million+ abstracts from the PubMed database are data mined by SemNet to create a knowledge graph, where 300,000 nodes are assigned node labels and each of the 20 million edges is assigned a relation type.
2. A Kaggle database that contains full-text COVID-19 articles is added to the existing knowledge graph. Next, all existing metapaths end up in the standard triplet format after being converted into heterogeneous embeddings.
3. GCN and GraphSAGE prediction methods are involved in training the model on triplet examples as well as negative examples.
4. TransE and RotatE are optimization methods used to score triplets utilizing negative sampling. An ensemble score using Borda count was preferable.
5. A web application was developed where arguments of any 2 entities are inputted, and a relation is predicted as a result.



Results

Query Node	Node Type	Degree
Anti-Inflammatory Agents	PharmacologicSubstance	3000
Antimalarials	PharmacologicSubstance	2028
Antiviral Agents	PharmacologicSubstance	2020
Envelope Proteins	AminoAcidPeptideOrProtein	276
Glycoproteins	AminoAcidPeptideOrProtein	14335
Immunomodulators	ImmunologicFactor	5105
Neuraminidase Inhibitors	PharmacologicSubstance	278
Nucleoside Analogs	PharmacologicSubstance	1365
Protease Inhibitor	BiologicallyActiveSubstance	6229

Table 1: UMLS Query Nodes used in the repurposed drugs simulations.

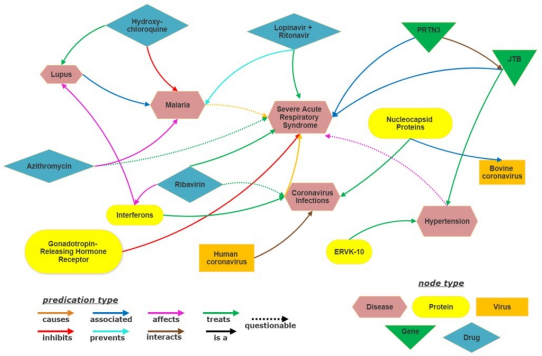


Figure 1: Pruned SARS knowledge graph. Nodes can be one of 132 types, and relationships can be one of 61 types. The actual knowledge graph generated consists of over 300,000 nodes but was downsized in this figure for simplicity.

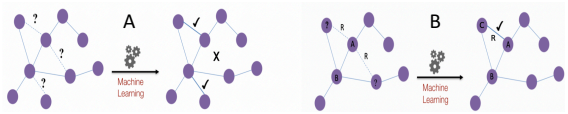


Figure 2: (A) The link prediction tool is used to confirm or deny the presence of relationships between nodes. (B) The link prediction tool is used to label both strong and questionable relationships between nodes. The labels can be of 61 types, encompassing many positive and negative relationships commonly found in medicine. Using **Figure 1** as an example, the link prediction tool could be used to predict a direct relationship between hydroxychloroquine and SARS, where currently no direct relationship exists.

Results

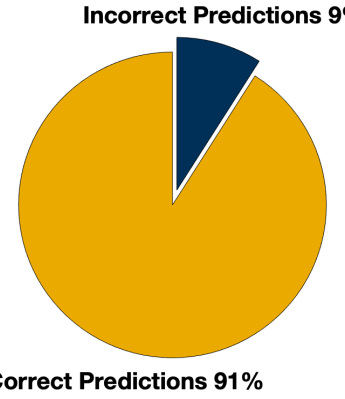


Figure 3: The model successfully predicted 91% of 44 peer reviewed, SARS coronavirus manually verified or "gold standard" relationships. These results provide insight into the possible efficacy of repurposed treatments for the novel COVID-19 virus.

Discussion

- In the 7 months since the global emergence of the COVID-19 pandemic, over 500,000 people have died worldwide, and the death count only continues to rise.
- In this work, several treatment options that had no established relation to COVID-19 in literature are now identified as potential treatments.
- In the future, we plan to validate the tool by comparing the results of the link prediction tool to that of recently published PubMed articles detailing new treatment options for COVID-19.

References & Acknowledgments

Acknowledgments:
Funding provided by the COVID-19 Whitaker Foundation pilot at Georgia Institute of Technology.

References:
1. Sedler AR, Mitchell CS. SemNet: Using Local Features to Navigate the Biomedical Concept Graph. Front Bioeng Biotechnol. 2019;7:1562.
2. Truchon M. Borda and the Maximum Likelihood Approach to Vote Aggregation. Mathematical Social Sciences. 2008;55(1):96–102