2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669736

COVID-19 Knowledge Graph for Drug and Vaccine Development

Lan Huang[a] , Hongrui Guan[a] , Yanchun Liang[b] , Renchu Guan[a*] and Xiaoyue Feng[a*]
[a] The Key Laboratory for Symbolic Computation and Knowledge Engineering of the Ministry of Education,
College of Computer Science and Technology,Jilin University Changchun,China
[b] Zhuhai Sub Laboratory, Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education,
Zhuhai College of Science and Technology,Zhuhai,China

Abstract—**The worldwide spread of COVID-19 has made a severe impact on human health and life. It has shown rapid propagation, long in vitro survival, and a long incubation period. More seriously, COVID-19 is more susceptible to variation, as it is an RNA virus. Mutations of COVID-19 have been reported in multiple countries worldwide, which makes drug and vaccine development a significant challenge. To search for potential drugs and vaccines and reveal the atlas of COVID-19 evolution, we extract information from massive unstructured data and construct a COVID-19 knowledge graph using the COVID-19 data. Based on machine learning approaches, we infer and predict novel coronavirus pneumonia-related diseases, drug action targets, etc. to speculate on new and more effective treatment methods. In addition, to study transcriptome of SARS-CoV-2, new ideas can be provided to biomedical experts with flexible responses to viral variation. An in-depth analysis of the COVID-19 pathomechanism at the pharmaceutical, genetic, and protein levels provides effective means and tools for novel coronavirus pneumonia vaccines, drug development, and therapeutic program design.**

Index Terms—COVID-19, knowledge graph, vaccine, mutation

I. Introduction

Since 2020, SARS-CoV-2 has undergone multiple mutations during its spread. There are currently four main variants. The Alpha variant appeared in England in September 2020. The beta variant appeared in South Africa in August 2020. The gamma variant was first discovered in the Amazon city of Manaus in December 2020. The Delta variant appeared in October 2020 and was first found in India. Each mutation brought great difficulties for the development of drugs and vaccines and posed a severe threat to the life and health of all humankind.

The randomness of gene mutations means that researchers in the medical field need to perform much clinical laboratory work to study the mutation of virus strains. Therefore, we hope that we can help medical experts reduce the manual workload, and in particular that we can provide help in the research and development of mutant strains and corresponding vaccine drugs.

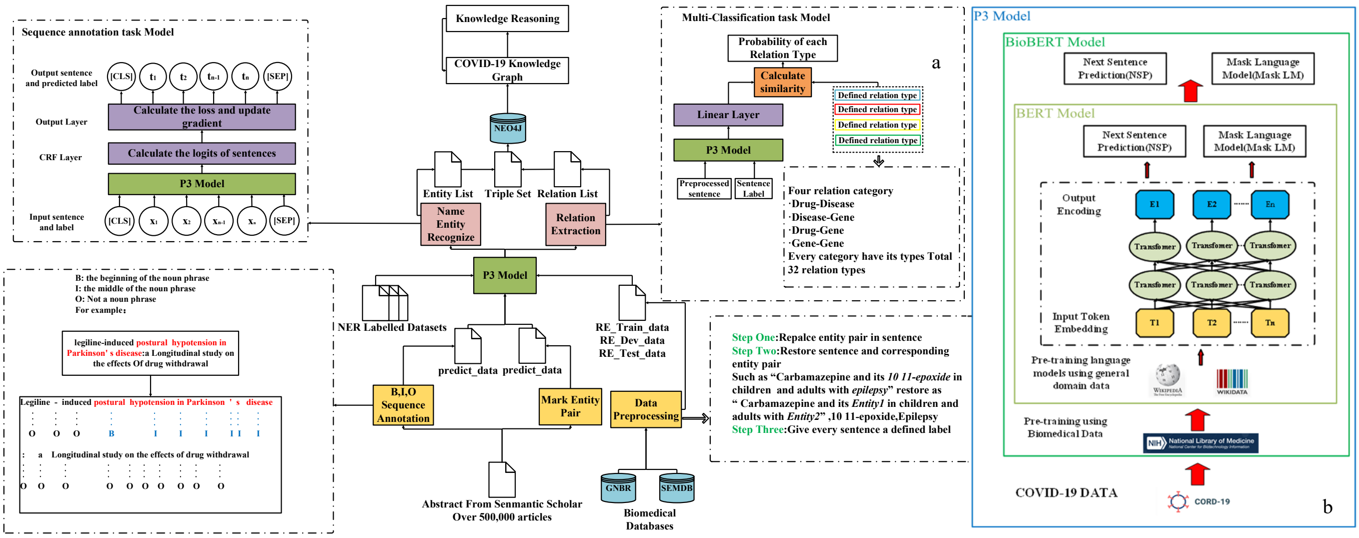
With this purpose, we start from a large amount of medical literature, including research literature on coronaviruses in the past few decades, such as SARS, and a large number of related studies since the outbreak of COVID-19. Data mining and natural language processing methods are used to realize the automatic construction of biomedical knowledge graphs. An automatically constructed biomedical knowledge graph can help medical researchers obtain the latest medical information more quickly, conveniently, and accurately. It can accurately store and query medical data to assist disease treatment and drug development.

More importantly, based on the biomedical knowledge graph, we can mine deeper information, such as the impact of genes and proteins on diseases, the corresponding targets of drugs, and the promotion or inhibition of drugs. Genetic information can help medical researchers find out possible genetic mutations, especially when research on a mutated strain of COVID-19 is imminent.

To construct a biomedical knowledge graph, we first extract entities and their relationships from scientific papers. This task can be divided into two parts, named entity recognition and relation extraction. Early research was conducted to complete the two tasks in a pipeline manner. In recent years, end-to-end method has been proposed to replace the pipeline model and most people believe that the end-to-end joint model can more accurately obtain the relationship between entities, and avoid the accumulation of errors caused by the pipeline model. However, Chen *et al.* [1] proposed a pipeline model and updated the state-of-the-art (SOTA) results in 2020. This made people rethink the choice between the joint model and the pipeline model. Therefore, we proposed a new model to achive SOTA results for COVID-19 knowledge discovery.

For biomedical data, it is more suitable to use two independent encoders for entities and relationships. Compared with general data, biomedical data are more difficult to model data mining tasks. It is because that biomedical data has more complex entity nesting relationships, and there is no unified entity representation. The specificity of entities is strong, and the relationship are complex and diverse. Therefore, the coding is necessary, which is more accurately corresponds to the entity extraction task and the relation extraction task. Even within each subtask, we may need to train models separately for different data to achieve the purpose of accurately mining data.

After an automated biomedical knowledge graph is initially constructed, the next step is to update the knowledge


Sequence annotation task Model
Output sentence and predicted label
[CLS]
t_1
t_2
t_{n-1}
t_n
[SEP]
Output Layer
Calculate the loss and update gradient
CRF Layer
Calculate the logits of sentences
P3 Model
Input sentence and label
x_1
x_2
x_{n-1}
x_n
Knowledge Reasoning
COVID-19 Knowledge Graph
NEO4J
Entity List
Triple Set
Relation List
Name Entity Recognize
Relation Extraction
P3 Model
NER Labelled Datasets
predict_data
predict_data
RE_Train_data
RE_Dev_data
RE_Test_data
B,I,O Sequence Annotation
Mark Entity Pair
Data Preprocessing
Abstract From Senmantic Scholar
Over 500,000 articles
GNBR
SEMDB
Biomedical Databases
B: the beginning of the noun phrase
I: the middle of the noun phrase
O: Not a noun phrase
For example：
legiline-induced postural hypotension in Parkinson' s disease:a Longitudinal study on the effects Of drug withdrawal
Legiline - induced postural hypotension in Parkinson ' s disease
O O O B I I I I I I
: a Longitudinal study on the effects of drug withdrawal
O O O O O O O O O O
Multi-Classification task Model
Probability of each Relation Type
Calculate similarity
Linear Layer
P3 Model
Preprocessed sentence
Sentence Label
Defined relation type
Defined relation type
Defined relation type
Defined relation type
a
P3 Model
BioBERT Model
Next Sentence Prediction(NSP)
Mask Language Model(Mask LM)
BERT Model
Next Sentence Prediction(NSP)
Mask Language Model(Mask LM)
Output Encoding
E1
E2
En
Transfomer
Input Token Embedding
T1
T2
Tn
Pre-training language models using general domain data
WIKIPEDIA
WIKIDATA
Pre-training using Biomedical Data
NIH National Library of Medicine
COVID-19 DATA
CORD-19
b

Four relation category
·Drug-Disease
·Disease-Gene
·Drug-Gene
·Gene-Gene
Every category have its types Total 32 relation types

Step One:Repalce entity pair in sentence
Step Two:Restore sentence and corresponding entity pair
Such as "Carbamazepine and its *10 11-epoxide* in children and adults with *epilepsy*" restore as " Carbamazepine and its *Entity1* in children and adults with *Entity2*" ,10 11-epoxide,Epilepsy
Step Three:Give every sentence a defined label

Fig. 1 COVID-19 Knowledge Graph and P3 Model. a. Flowchart of the COVID-19 Knowledge Graph Model; b. The framework of P3 model

graph and discover knowledge based on the knowledge graph. From the beginning of data preprocessing to entity relationship extraction, our model can be completed automatically. Therefore, as long as the latest medical literature is obtained, the information of the literature can be added to the knowledge graph.

Finally, we use knowledge reasoning methods to predict entities that potentially have relationships with COVID-19 and discover new relationships. In addition, we use the time-slicing method to verify the accuracy of the knowledge inference results. We can intercept the data for a certain period, make predictions based on the data, and then compare the results with the corresponding medical research results.

II. Method

Our framework is divided into two main parts. One is the automatic construction of the COVID-19 knowledge graph. The second is knowledge reasoning based on the knowledge graph. The main task of constructing a domain knowledge graph is information extraction, including named entity recognition (NER) and relationship extraction (RE). The entities and their relationships are extracted from massive unstructured scientific papers to generate "entity-relation-entity" triplets. The types of entities and relationships determine the knowledge graph's complexity and indicate the quality of the information contained in the knowledge graph. Our COVID-19 knowledge graph includes five classes of entities: drugs, diseases, genes, proteins, and compounds. Six relationships are included: disease–drug interaction (DDI), disease–gene interaction (DGI), disease–protein interaction (DPI), gene-protein interaction (GPI), compound-disease interaction (CDI) and disease–chemical interaction (DCI).

A. Data preprocessing

Biomedical data are intricate and cannot be used directly. Therefore, we preprocess the datasets and transform the data into formats that can be applied to machine learning without undermining the medical value of the data. We introduce the data preprocessing operation separately for entity abstraction and relation abstraction task.

First, we use the abstracts of COVID-19 scientific literatures from academic websites for sentence and word segmentation work. We introduce the Stanford word splitting tool standoff2conll when performing sequence annotation work and give every input word an annotation of "O", as shown in Fig. 1(a), for the sequence annotation task.

The second module is the data annotation of the training set versus the validation set for the relationship extraction task. We obtain COVID-19-related biomedical data from GNBR, including drugs, diseases, genes, proteins, compounds, and the relationships among them. Then, we compose the above relations with the entity pairs into the corresponding "entity-relation-entity" triple set. In the next step, we will label the entity information in the triad, including the entity's properties and where the entity is located in the original text. Annotated relation types of text are used for the training and validation sets in classification tasks. At the same time, we need to label the pairs of entities. Our classification model finally predicts the relationships between pairs of entities.

B. Named Entity Recognition

In our framework, we take the entity recognition task as a sequence labeling process. We use the BIO labeling method to assign the label "O" to each input token and use statistical machine learning methods to predict sequence labels. We use the conditional random field (CRF) [2] as a probabilistic graph model. The CRF model can find the optimal sequence that maximizes the objective function and complete the task of named entity recognition. The input sequence is $X = (x_1, x_2, \ldots, x_n)$, and the output sequence is $Y = (y_1, y_2, \ldots, y_n)$. The transfer matrix of CRF is A. A_{ij} represents the transition probability from y_i to y_j.The start and end tags of the sentence are included. P represents the non-normalized probability of X mapping to Y We define the score function as Eq. (1):

$$\mathrm{S(X,y)} = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \quad (1)$$

We use the softmax function to define a probability value for each correct label sequence, where Y_X are all possible predicted label sequences:

$$p(\mathrm{y}|\mathrm{X}) = \frac{e^{s(\mathrm{X},\mathrm{y})}}{\sum_{\tilde{y}\in Y_X} e^{s(\bar{X},\tilde{y})}} \tag{2}$$

During training, we maximize the log likelihood of $p(y|X)$:

$$\log(p\big((y|X)\big) = s(X,y) - \log\big(\textstyle\sum_{\tilde{y}\in Y_X} e^{s(X,\tilde{y})}\big) \tag{3}$$

The training loss function is given by Eq. (4):

$$\mathrm{L} = -\log\big(p(\mathrm{y}|\mathrm{X})\big) = \log\big(\textstyle\sum_{\tilde{y}\in Y_X} e^{s(\bar{X},\tilde{y})}\big) - \mathrm{s}(\mathrm{X},\mathrm{y}) \tag{4}$$

C. Relation Extraction

For the input tokens X = (***x₁, x₂, ... xₙ***), we define the span to represent the entity extracted by NER (some biomedical entities consist of multiple words, so we define the span to represent them). The task of relation extraction is to select a relation from the relation sets *R* for each span pair. At the same time, the relationship type set contains "NONE", that is, there is no relationship between span pairs. For every span *i*, the representations of entities and relations are:

$$R_e(s_i) = (X_{START(i)}, X_{END(i)}) \tag{5}$$

$$R_r(s_i, s_j) = \big(\hat{X}_{\widehat{START}(i)} \ldots \hat{X}_{\widehat{END}(i)}\big); \big(\hat{X}_{\widehat{START}(j)} \ldots \hat{X}_{\widehat{END}(j)}\big) \tag{6}$$

We take the vector representation corresponding to the last layer of the [CLS] mark. [CLS] is the abbreviation of classification, which means that the model is used for downstream text classification tasks. For single-text classification tasks, the [CLS] mark indicates the starting position of the input text. For sentence pair classification tasks, [CLS] appears at the beginning of the sentence pair, and together with the [SEP] (Sentence separator) identifier at the end of each sentence marks the two sentences of the sentence pair. Then linear transformation and softmax normalization are performed to obtain the classification probability. During fine-tuning, all parameters of the BERT model and task-related layers are updated together to optimize the loss function.

In the process of relationship extraction, we use the softmax function to define a probability value for each correct relationship label:

$$P_r(\mathrm{r}|s_i, s_j) = Softmax\Big(W_r R_r(s_{i,} s_j)\Big) \tag{7}$$

TABLE I. THE COMPRESSION OF THE PERFORMANCE AMONG DIFFERENT PRETRAINED MODELS AND OUR MODEL

	BioBERT	**XLNet**	**RoBERTa**	**ERNIE**	**Our model**
F1 Score	83.74	79.11	73.36	63.38	**84.82**
Recall	90.75	88.57	79.64	61.77	**91.79**
Precision	77.74	71.47	67.99	65.08	**78.83**
Specificity	71.15	61.02	58.66	62.14	**72.83**

D. Multitask learning

Among the many information extraction models, the multitask learning model optimizes the entity extraction and relationship extraction tasks at the same time through parameter sharing. Multiple related tasks are combined and multiple tasks are learned simultanouesly. Some factors are shared among the tasks. The information learned in the learning process can be shared, which is also a feature of single-task learning. Associated multitask learning can achieve better generalization effects than single-task learning.

Our model is of a single pipeline and the fine-tuning completes only one downstream task at one time. However, in the relation extraction part, we use shallow sharing (shared representation) to have the classifiers share and complement each other's learned domain information (domain information), promote each other's learning ability, and enhance the effect of generalization.

E. Pretrained language model

The one-hot encoding generates a high-dimensional and sparse matrix that is accompanied by the "curse of dimensionality". In 2003, Bengio *et al.* [3] proposed Nerual Network Language Model (NNLM), which began the era of neural network models. The emergence of the pretrained anguage model means that the model parameters are no longer randomly initialized. The pretrained language model obtains the model parameters through pretraining on specific tasks. Although word vector models such as Word2vec [4] and GloVe [5] improve the efficiency of word vector generation, they cannot determine the polysemous and complex semantics of a word. Recently, pre-trained models such as ELMo [6] and BERT [7] have achieved SOTA results in natural language processing (NLP) tasks. The full name of BERT is Bidirectional Encoder Representation from Transformers, which is the encoder of two-way transformer. BERT builds a language model by using Mask Language Model (MLM) pre-training tasks. In the training process, 15% of the tokens are randomly masked, then these tokens are predicted. The MLM task is mainly used for text classification tasks, which is one of the NLU (Natural Language Understanding) task. After BERT was proposed, some pre-trained language models were successively proposed and SOTA was achieved on some tasks, such as UNILM [8], ERNIE [9], Roberta [10], XLNet [11], and BioBERT [12]. The overall framework of BioBERT is shown in the Fig. 1(b). The results of the comparative experiment are shown in Table I.

F. Knowledge Graph Embedding and Reasoning

Large-scale knowledge graphs that have been successfully applied to many scenarios, such as Freebase [13], DBpedia [14], YAGO [15], and NELL [16], are still incomplete. Therefore, the embedding and completion of knowledge graphs are important follow-up tasks for knowledge discovery. Knowledge graph embedding methods can be divided into translation distance models and semantic matching models. The former models utilize distance-based scoring functions. Examples include TransE[17], TransH[18], TransR [19], TransD[20] and RotatE[21]. They use the distance between two entities to measure the rationality of a fact. The semantic matching model uses a similarity-based scoring function. For example, DisMult[22] measures the credibility of facts by matching the latent semantics of entities and relationships.

G. Our Model

In the BioBERT model, relationship abstraction is defined as a binary classification task. Therefore, to accurately complete data mining and obtain more valuable forecast information, we redefined the multiple classification model.

We used COVID-19 data to pretrain the BioBERT model, making the model more suitable for the data mining of COVID-19. We call our model P3 language model (three-stage pre-training). The framework of P3 model is shown in the Fig. 1(b).

III. Experiment

A. Experimental environment

Our experimental platform is Linux and the programming language is Python. We use four Tesla V100 GPUs to accelerate calculations.

B. Pretraining BioBERT

To improve the performance of the pretrained language model, we use the COVID-19 literature as pretraining data and erform pretraining based on BioBERT. Using this method, the training language model is more suitable for COVID-19 research and knowledge discovery, such as obtainingmore accurate COVID-19 entity relationships.

C. Dataset

Entity extraction dataset. To improve the quality of the entity extraction results, we conducted experiments on multiple datasets and fused the results of each dataset. The annotation sets used in our experiments were NCBI, BC5CDR and JNLPBA.

Relation extraction dataset. The relation extraction data come from GNBR [23]. With the data preprocessing, we obtain the classification annotation. As shown in Fig. 1(a), the annotated dataset contains four kinds of relationships. Each relationship type is further divided into multiple subtypes. There are 32 relationships in total.

COVID-19 dataset. Our data comes from [24], 293,599 articles in the medical field related to COVID-19 collected from March 2020 to October 2020 as the start-up data. Meanwhile, with the weekly updates, we are constantly expanding the data set to increase the scale of the knowledge graph. Till to writing this paper, we extracted knowledge from over 500,000 articles.

D. Metrics

We choose accuracy and specificity as the evaluation indexes. Specificity refers to the proportion of the predicted value of the model in all the results whose real value is negative. Compared with accuracy, it is more significant for natural language processing models in medical field. From the perspective of security, more attention is given to the mis-information identification. The formula of specificity is given by Eq. (8). *TP* is true positive, indicating a positive example that has been correctly classified; *FN* is false negative, indicating that a positive example is mistakenly classified as negative.

$$specificity = \frac{TP}{TP+FN} \tag{8}$$

IV. Results

A. Results of Pretraining

Based on BioBERT, P3 model is retrained with COVID-19 literature data with seven days. Fig. 2 shows the results of the pretrained model with different epochs.

B. Result of NER

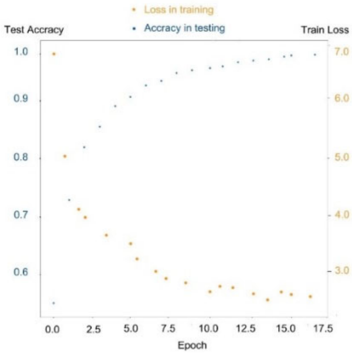


Fig. 2 Result of pre-trained P3 model

TABLE II. The compression results of named entity recognition between our model and BioBERT

	OUR MODEL			**BioBERT**		
	Pre-cision	***Recall***	***F1***	***Pre-cision***	***Recall***	***F1***
BC5CDR CHEM	0.908	**0.925**	0.916	0.9426	0.923	0.933
BC5CDR DISEASE	0.837	**0.848**	0.843	0.8961	0.831	0.862
JNLP	**0.897**	**0.917**	**0.907**	0.7443	0.832	0.786
NCBI	**0.909**	**0.904**	**0.906**	0.8830	0.890	0.886

As shown in TABLE II, the language model pretrained on COVID-19 data achieved better performance than BioBERT in the downstream task of NER. In particular, the improvement effect is evident on the JNLP dataset. We believe that in named entity recognition, the reason for the apparent improvement in the effect of gene type entity extraction is that the pretraining corpus contains more data about genes. This enhances the ability of language models to understand genetic data.

C. Result of Relation Extraction

In the relationship extraction task, we conducted three types of relationship prediction experiments, CGI, DDI and GPI. The results are shown in TABLE III.

TABLE III. Relationship prediction results

Relation Type	**Specificity**
Compound-Disease-Interaction (CGI)	0.5801
Disease-Drug-Interaction (DDI)	0.5023
Gene-Protein-Interaction (GPI)	0.4061

D. Extracted Knowledge

With entity recognition, we obtained 10,993 biomedical entities from the literature, including the five categories of drugs, diseases, genes, proteins, and compounds. Combining the results of relationship extraction, we extracted 1,204,234 triples. In addition, we integrated some of the relationships in the SEMDB database. The fusion criterion is as follows: if entity A appears in both SEMDB and our extracted entity set simultaneously, we add the entity A-related relationship from the SEMDB database to our relationship set to expand the knowledge graph. We assign an ID and LABEL to all entities and relationships and save all entity and relationship triples.

E. Results of Knowledge Graph Embedding and Completion

As shown in TABLE IV, our knowledge graph achieved good results in both the distance-based representation model and the semantic-based representation model. We found that the RotatE-based models performed the best. This is because our knowledge graph has rich relationships. The relationships between biomedical entities are more complicated. For example, different dosages of drugs have different effects on diseases, and a pair of drugs and targets of action will also exhibit other interaction relationships under different conditions. Therefore, the types of relationships in the graph are more complicated. This is the reason why the TransE model performs poorly on our data.

TABLE IV. THE RESULTS OF THE GRAPH REPRESENTATION OF THE COVID-19 KNOWLEDGE GRAPH

	HITS@1	HITS@3	HITS@10
pRotatE	**0.414**	0.423	0.437
TransE	0.134	0.378	0.456
RotatE	0.401	**0.443**	**0.491**
DisMult	0.347	0.381	0.413
ComplEx	0.345	0.381	0.418

TABLE V. PREDICTED ENTITIES ASSOCIATED WITH COVID-19

Name of genes or Proteins	Verified
IL-6	**Yes**
ACE2 gene	**Yes**
SARS 3CL protease	**Yes**
6-phosphogluconate dehydrogenase	No
IL-1 RA	No
CXCL10 mRNA	No
pIFN-γ	No
Human GM-CSF	No
CCR2 genes	No
MHC-I	**Yes**
IDR	No
carboxypeptidase	No
TGF-β1	No
cTn	**Yes**
sPLA2	No
OAS	**Yes**

Since the mutations of COVID-19 are closely related to genes and proteins, as shown in TABLE V, we selected the top 20 genes and proteins with the highest knowledge completion rankings, which were the most relevant to COVID-19 in our knowledge graph. The data we used are from March to October 2020. Some of the genes and proteins (Labeled Yes in Table V) we predicted have been confirmed to be closely related to COVID-19 mutation and vaccine development. The others (Labeled No in Table V) are waiting to be verified.

V. ANALYSIS

Due to the single-stranded structure of COVID-19, the virus produces multiple mutant strains, which are considerable obstacles for vaccine development. Therefore, we extracted the novel coronavirus pneumonia-related networks from the COVID-19 knowledge graph. They can help us find the latest and most effective treatment methods for novel coronavirus pneumonia and promote the development of vaccines.

In addition to knowledge reasoning based on the graph representation learning, we used the path discovery method to discover the relationship between COVID-19 and genes or proteins. The extracted paths can be seen as proof from the knowledge reasoning result. In the knowledge graph stored in the Neo4J, we searched for the paths with 1 to 5 hops from COVID-19 to its related genes or proteins. Some of the paths are shown in Fig. 3.

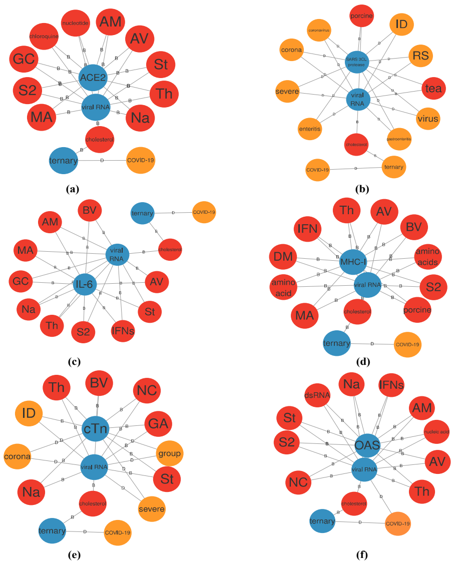


Fig. 3 The paths are verified to be related to COVID-19

ACE2. Fig. 3(a) shows the path from COVID-19 to the ACE2. Angiotensin-converting enzyme 2 (ACE2) is a protein on the surface of cell membranes. Many studies have shown that ACE2 is a crucial receptor allowing SARS-CoV-2 to invade human cells. In addition, scientists have found in mouse models, there is a positive correlation between the expression of hACE2 and the degree of SARS-CoV-2 infection [25].

3CLPro. Fig. 3(b) shows the path from COVID-19 to SARS 3CLpro. Medical experts have studied the affinity of several anti-HIV-1 drugs with SARS-CoV main protease. The results show that darunavir has the best binding affinity. This provides new paths for the development of anti-COVID-19 clinical drugs. The structural design of anti-COVID-19 drugs targeting the SARS 3CLpro was clarified in [26].

IL-6. Fig. 3(c) shows the path from COVID-19 to IL-6. IL-2, IL-4, TNF-α, IFN-γ and CRP are not correlated with the severity of COVID-19, while only IL-6 and IL-10 are positively correlated with the severity of COVID-19. It was independently discovered by Harbin Medical University, Harbin Veterinary Research Institute, the State Key Laboratory of Veterinary Biotechnology, Wuhan University People's Hospital, and Wuhan University's State Key Laboratory of Virology [27]. IL-6 and IL-10 can be used as the markers to monitor the condition of patients with severe COVID-19.

MHC-I. Fig. 3(d) shows the path from COVID-19 to MHC-I. Studies have shown that some variants of MHC-I can protect the body by stimulating a robust immune response. Nevertheless, some may make individuals vulnerable to viral attacks, serious illnesses, and death. A new study tested 52 MHC-I alleles and found significant differences in their ability to combine segments from the entire genome of the novel coronavirus and essential protein segments that generate a robust immune response [28]. Novel coronavirus pneumonia data from 23 countries show that the mortality of this disease is closely related to the distribution of MHC-I variants.

cTn. In a study of 41 patients published in the medical journal The Lancet on January 24, there were 5 cases of acute heart injury, accounting for 12% of the patients, and these 5 cases had a significant increase in troponin I (cTnI) [29]. In a study of 99 patients published in The Lancet, 15 points (15%) of myoglobin were elevated. In Fig. 3(e), we not only found a path from cTn to COVID-19 but also accurately predicted the direct relationship between cTn and COVID-19.

Oligoadenylate synthetase (OAS) can encode a protein that activates enzymes that break down viral RNA. Any of these genetic changes may change this code, which allows the virus to multiply. Research data show that this mutation, like the genetic risk factor for interferon, will impact the progression of novel coronavirus pneumonia. Fig. 3(f) shows the path from COVID-19 to OAS.

VI. Conclusion

In this article, the COVID-19 data used in our knowledge graph are the literatures up to October 2020. The information on the above six paths was confirmed by the research after October 2020. It shows that these results have good credibility and research value. We verified some of the predicted relationships are correct using the time-slicing method. The remaining unverified ones are waiting for the follow-up researches. Due to the automatic construction mechanism of the knowledge graph, we can obtain more medical information about COVID-19 based on the latest research literatures and predict more biomedical knowledge.

Acknowledgment

The authors are grateful for the support of the National Key Research and Development Program of China (2018YFC20013-02), National Natural Science Foundation of China (61972174, 62172187), Science and Technology Development Plan Project of Jilin Province (20200403172SF, 20200801033GH, YDZJ-202101ZYTS128), Science and Technology Planning Project of Guangdong Province (2020A0505100018), Guangdong Key Project for Applied Fundamental Research (2018KZDXM076) and Bioknow MedAI Institute (BMCPP-2018-002).

References

[1] Z. Zhong, D. Chen,"A Frustratingly Easy Approach for Entity and Relation Extraction," in Proceedings of NAACL-HLT, 2021, pp. 50-61.

[2] J. Lafferty, A. McCallum and FCN. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in Proceedings of ICML, 2001, pp. 282-289.

[3] Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, "A neural probabilistic language model," The journal of machine learning research 3, 2003, pp. 1137-1155.

[4] Y. Goldberg, O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," arXiv preprint arXiv:1402.3722, 2014.

[5] J. Pennington, R. Socher, CD. Manning, "Glove: Global vectors for word representation," in Proceedings of EMNLP, 2014, pp. 1532-1543.

[6] ME. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in Proceedings of NAACL-HLT, 2018, pp. 2227-2237.

[7] J. Devlin, MW. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171-4186.

[8] D. Li, et al, "Unified language model pre-training for natural language understanding and generation," in Proceedings of NIPS, 2019, pp. 13063-13075.

[9] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, "ERNIE: Enhanced Language Representation with Informative Entities," in Proceedings of ACL, 2019, pp. 1441-1451.

[10] Y. Liu, et al, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[11] Z. Yang, Z. Dai, Y. Yang, et al, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in Proceedings of NIPS, vol. 32, 2019, pp. 5753-5763.

[12] J. Lee, et al, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics vol. 36, 2020, pp. 1234-1240.

[13] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in Proceedings of SIGMOD, 2008, pp. 1247-1250.

[14] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in the semantic web, Berlin, Heidelberg, 2007, pp. 722-735. Springer.

[15] FM. Suchanek, G. Kasneci and G. Weikum, "Yago: a core of semantic knowledge," in Proceedings of WWW, 2007, pp. 697-706.

[16] C. Andrew, et al, "Toward an architecture for never-ending language learning," in Proceedings of AAAI, 2010, pp. 1306–1313.

[17] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in Proceedings of NIPS, vol. 26,2013, pp. 1-9.

[18] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in Proceedings of AAAI, vol. 28, no. 1, 2014, pp. 1112-1119.

[19] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in Proceedings of AAAI, 2015, pp. 2181-2187.

[20] G. Ji, S. He, L. Xu, K. Liu, J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in Proceedings of ACL, vol. 1, 2015, pp. 687-696.

[21] Z. Sun, ZH. Deng, JY. Nie, J. Tang, "RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space," in Proceedings of ICLR (Poster), 2019.

[22] B. Yang, W. Yih, X. He, J. Gao, L. Deng, "Embedding Entities and Relations for Learning and Inference in Knowledge Bases," in Proceedings of ICLR (Poster) , 2015.

[23] B. Percha, RB. Altman, "A global network of biomedical relationships derived from text," Bioinformatics, vol. 34, no, 15, 2018, pp. 2614-2624.

[24] LL. Wang, et al, "CORD-19: The COVID-19 Open Research Dataset," in Proceedings of ACL 2020, 2020.

[25] L. Gaziano, et al, "Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19," Nature medicine, vol. 27, no. 4, 2021, pp. 668-676.

[26] MT. ul Qamar, et al, "Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants," in Journal of pharmaceutical analysis, vol. 10, no.4, 2020, pp. 313-319.

[27] H. Han, et al, "Profiling serum cytokines in COVID-19 patients reveals IL-6 and IL-10 are disease severity predictors," in Emerging microbes & infections, vol. 9, no. 1, 2020, pp. 1123-1130.

[28] EA. Wilson, et al, "Total predicted MHC-I epitope load is inversely associated with population mortality from SARS-CoV-2," in Cell Reports Medicine, vol. 2, no. 3, 2021, pp. 100221.

[29] C. Huang, et al, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," in The lancet, vol. 395, no. 10223, 2020, pp. 497-506.