Drug Targeting for COVID-19 by Comparing It to Other Viral Diseases with a Computational Approach

Ju Hyun Kim≠, Hye Won Oh≠, Yu Ri Kim≠, So Yon Park≠, Wan Kyu Kim*

Department of Life Science, Ewha Womans University, Seoul 03760, Korea

≠ These authors contributed equally to this work

Abstract

The total number of COVID-19 cases has reached about 200 million people worldwide. Despite the vaccination, Korea is reporting more than 1,000 new confirmed daily cases for over a month from July 2021. To alleviate this situation, it is necessary to provide treatment, not just vaccination. Therefore, this study aims to find a candidate drug that will be a treatment for COVID-19 by analyzing the treatment drugs for influenza and MERS, which are similar to COVID-19. We firstly present a differentially expressed gene (DEG) in the drug treatment experimental data published by GEO. In addition, we obtained a list of DEGs that are common in disease data that will serve as a comparison group and RNA sequencing data of COVID-19. Next, the two DEG lists obtained previously were compared and the relationship between the pathways involved in the DEG was analyzed. Following the results of analysis obtained through the study, we predicted the possibility of the COVID-19 treatment applying each drug into the CNN model trained as COVID-19 related compounds according to their molecular formula similarity. As a result, we suggest that an in-silico method derives COVID-19 candidate drugs more efficiently by cross-verifying them into two ways.

Methods

Data preprocessing to get Raw counts by RNA sequencing

Data relative to the transcriptional response to SARS-CoV-2 infection and healthy controls were obtained from datasets deposited within the Gene Expression Omnibus (GEO) (GSE160351) using SRA-Tools (version 2.9.2).[1] In the same way, influenza (GSE154596, and GSE93999) and MERS (GSE122876) data, including samples treated with specific drugs, were acquired. As a result, 3 samples of COVID-19 patients (3 controls), 5 samples of influenza patients (5 controls, and 4 treated samples), and three MERS patient samples (each 3 controls, and treated samples) were used in the experiment.

RNA-Seq pipeline

First, all raw RNA-seq reads obtained by GEO were quality-checked using the software FASTQC (version 0.11.8). Next, FASTP (version 0.21.0) was used for trimming paired-end reads and BBduk (version 38.50b) for single-end reads. Then, raw read counts were obtained by HTSeq, after trimmed data were aligned with STAR (version 2.7.1) on the GRCh38 genome guided by GENCODE annotation (version 34). [1]

Identification of Differentially Expressed Genes and Pathway Analysis

The results of extracting only protein coding genes from all raw counts data were filtered and normalized through the R package, 'edgeR' (version 3.28.1). Also, quasi-likelihood F-tests provided by the edgeR package were conducted to find differentially expressed genes (DEGs). To find the pathways involving DEGs, the 'goseq' package (version 1.44.0) was used to perform Gene Ontology (GO) enrichment analysis, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were identified with the 'clusterProfiler' package (3.14.3).

SMILES notation

Simplified molecular-input line-entry system (SMILES) is a chemical notation language devised by David Weininger in 1989 [2] which enables molecular machine computation for scientists. SMILES uses a set of atomic symbols and SMILES original symbols. In SMILES representation, instead of structural formula, atoms are represented by their atomic symbols. For example, double bonds are written using "=" and triple bonds using "#". Rings are represented by breaking one bond per ring in a connected molecule and the presence of the ring is indicated by appending an integer to each of the two atoms of the broken bond. The branches are indicated by balanced pairs of parentheses, an "(" is followed at some point by a corresponding ")". It means that all the atoms in that branch have included a chain. For instance, the SMILES string for Acetaminophen is CC(=O)NC1=CC=C(C=C1)O. To conduct this experiment, we utilized a normalization algorithm to ensure the one single SMILES representation (canonical SMILES) from one compound[3].

Convolution Neural Network (CNN) and SMILES convolution fingerprint model (SCFP)

CNNs allow learning data-driven, highly representative, hierarchical image features from sufficient training data.[4] Convolutional neural networks are a specialized type of neural network that uses convolution different than general matrix multiplication in at least one of their layers.

SMILES convolution fingerprint model transforms the SMILES feature matrix into a low-dimensional feature vector. The SMILES string, which is an input used for the CNN model, is required to be converted into numeric forms to calculate. Therefore, for each symbol in the SMILES string, a feature vector that is a distributed representation of the symbol is calculated as one-hot vectors for each. For this SCFP model, 42 features are contained, which 21 features used as symbols for atoms and the remaining 21 features for the original SMILES symbols. In this way, a low-dimensional feature vector transformed by CNN termed the SMILES convolution fingerprint (SCFP)[5]. This SCFP model allows the SMILES feature matrix to be applied in a similar way to the conventional CNNs to image dataset with the highest ROC-AUC value when it compares with the other models as TOX21 datasets[5].

For the fitness of our dataset, we made some extent changes for the original SCFP model and reflect the package module updates(www.dna.bio.keio.ac.jp/smiles). We applied this tuned SCFP model to compare two drugs that were treated on MERS and Influenza for each in previous experiments. To maximize the learning effect for our datasets by adjusting the learning rate as stable, we changed the optimizer of the original from Adam into SGD [6].

Preparation of the Dataset and Evaluation

In this study, we used the datasets from PubChem, the severe acute respiratory syndrome coronavirus 2(SARS-CoV2) datasets. The raw dataset contains information about whether approximately 26740 chemical compounds would relate to SARS-CoV2. Among the chemical compounds included in this dataset, only active and inactive notated chemical compounds were labeled 0 and 1, respectively, and used as the experimental dataset. In addition, a thousand labeled chemical data were changed into canonical SMILES notation each and divided this dataset with a ratio of train:test=9:1. We evaluated the performance of the SCFP model by using the area under the ROC-AUC, which is widely used for measurement to evaluate the performance of classifiers. The ROC-AUC has a value from 0 to 1, where a higher value indicates a more accurate classification between active and inactive compounds.

Results

Identifying the target pathway of each treatment

The data used in this study were extracted under different conditions (cell line, cell type, etc.) in different laboratories for each disease. Therefore, direct comparison between disease data is not available, so a method of comparing the results of analyzing the pathways involved in the disease was adopted.

Selection of Differentially Expressed Genes

In the COVID-19 data, three control and three disease samples were compared for finding the genes that were higher or lower expressed than normal control samples by applying a result of quasi-likelihood F-tests, FDR < 0.05 & |logFC| > 1.5. As a result, 142 up DEGs and 571 down DEGs were selected.

Data for Influenza experiment(GSE93999) involved 1 control, 1 untreated disease, 2 low concentration of drug-treated samples (400 ug/ml), and 2 high concentration of drug-treated samples (800 ug/ml). Since this data lacked replicates, analysis was performed after adding another RNA-seq data from the same cell line for more accurate experiment by correcting errors. A total of three different DEG analyses were performed for comparison between samples of Influenza. First, DEGs (up: 145, down: 1383) were identified between the control and the untreated disease samples. Another set of DEGs were confirmed by comparing the disease samples that were not treated with the drug and the disease samples treated with the Lariciresinol-4-β-D-glucopyranoside by concentration (low and high). As a result, the number of up DEGs (1519 for low, and 1391 for high dose) and down DEGs (723 for low, and 333 for high dose) of the treated samples were confirmed only when FDR < 0.05 & |logFC| > 2 were satisfied.

Similarly, in the MERS data compared between control and disease samples (also, between untreated disease and AM580-treated disease samples), the number of overexpressed genes (263 and 570, respectively) and the number of underexpressed genes (613 and 286, respectively) in disease samples versus control (and disease samples treated with drugs versus disease samples without drug) were identified by applying criteria, FDR < 0.01 & |logFC| > 2.

Disease	Criteria	UP	DOWN
COVID-19	FDR < 0.05 & |logFC| > 1.5	142	571
Influenza (control vs. untreated)	FDR < 0.05 & |logFC| > 2	145	1383
Influenza (untreated vs. treated -low-)	FDR < 0.05 & |logFC| > 2	1519	723
Influenza (untreated vs. treated -high-)	FDR < 0.05 & |logFC| > 2	1391	333
MERS (control vs. untreated)	FDR < 0.01 & |logFC| > 2	265	613
MERS (untreated vs. treated)	FDR < 0.01 & |logFC| > 2	570	286

[table1 Criteria for DEG selection]

Finding common DEGs

In order to find the target pathway of the two types of treatment used in the influenza and MERS data, the common genes between the overexpressed genes in disease samples (compared to control samples) and the genes that were underexpressed in drug-treated disease samples (compared to untreated disease samples) were extracted. Similarly, the common genes between down DEGs of disease samples and up DEGs of treated samples were also identified.

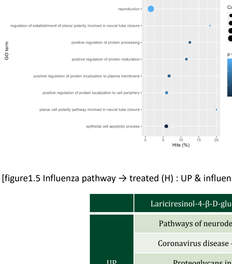
As a result, the top10 pathways of Gene Ontology (GO) enrichment analysis and the results of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analysis were obtained on the basis that the pathway in which the common gene sets were involved becomes the target pathway for each treatment.

[figure1.1 MERS pathway → treated : UP & MERS : down]

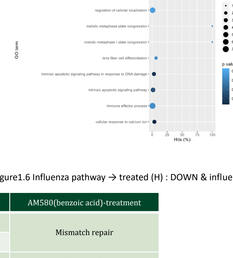
[figure1.2 MERS pathway → treated : DOWN & MERS : up]

[figure1.3 Influenza pathway → treated (L) : UP & influenza : down]

[figure1.4 Influenza pathway → treated (L) : DOWN & influenza : up]



[figure1.5 Influenza pathway → treated (H) : UP & influenza : down]



[figure1.6 Influenza pathway → treated (H) : DOWN & influenza :up]

	Lariciresinol-4-β-D-glucopyranoside	AM580(benzoic acid)-treatment
UP	Pathways of neurodegeneration	Mismatch repair
UP	Coronavirus disease – COVID-19	
UP	Proteoglycans in cancer	Ferroptosis
UP	Notch signaling pathway	Thiamine metabolism
UP	Arginine biosynthesis	
DOWN	PI3K-Akt signaling pathway	Non-alcoholic fatty liver disease
DOWN		complement and coagulation cascades

[table2 Summary of KEGG pathway analysis]

However, it was difficult to identify a direct relationship between the main pathway of COVID-19 and the target pathway involved in the treatments of influenza and MERS. Still, it was confirmed that COVID-19 was included in the KEGG pathway of AM580 treatment, even if the p-value was not significant. Therefore, in order to find commonalities between target pathways of each treatment and COVID-19 pathway, we tried to confirm whether the two treatments regulate pathways related to COVID-19, through other literature or research results.

Pathways associated with SARS-CoV-2 and potential therapeutic targets

We previously classified several pathways expected to have a significant association with the COVID-19 through the analysis of influenza and MERS drug treatment DEG. Since then, we have tried to determine which of them is actually noteworthy as a "potential therapeutic target" of the COVID-19. According to what has been studied so far, 14 of the approximately 60 pathways we have identified actually have a significant association with the COVID-19 and concluded that they are noteworthy in making treatments for the COVID-19 in the future.

Neurodegeneration of MERS pathway was first noted. Since neuroinflammation can be initiated or deteriorated by the infection contamination itself, just as by upsetting conditions like those connected to the new pandemic, the job of neuroinflammatory instruments could be focal in an endless loop prompting an expansion in the mortality hazard in matured COVID-19 patients. Moreover, set off neuroinflammatory pathways and resulting neurodegenerative and neuropsychiatric conditions may be possible long haul confusions of COVID-19.[7] Some researcher observed that the metabolic pathways of VA against CHOL/COVID-19 were involved in arginine biosynthesis; glyoxylate and dicarboxylate metabolism; alanine, aspartate, and glutamate metabolism; arginine and proline metabolism, and tryptophan metabolism.[8] One study mentioned the effect of ferroptosis on SARS-CoV-2 infection. According this study, uncovering the role of ferroptosis in SARS-CoV-2 infection is essential to develop new treatment strategies for COVID-19. Intracellular cell iron depletion or new generation of ferroptosis inhibitors might be potential drug candidates for COVID-19.[9] Thiamine metabolism pathway especially has been mentioned repeatedly, in several studies, suggesting a deep association with COVID-19. Thiamine use as adjunctive treatment might have potential endurance benefits in fundamentally sick patients with COVID-19. Moreover, it was related with a lower rate of apoplexy. Further interventional studies are needed to affirm these discoveries.[10][11]

Recently, activating the CAP has also been suggested a therapeutic strategy for respiratory diseases. And cholinergic synapse pathway influencing circulating TNF amounts and the shock response to endotoxaemia is likely to be a hopeful therapeutic intervention in COVID-19 infection.[12]

Furthermore, common inflammatory pathways associated with COVID-19 pathophysiology and GLP-1 functions have been discussed worldwide, in order to outline the anti-inflammatory effects of GLP-1 in different systems of the body and evaluate its potential applications in the diagnosis and treatment of patients with COVID-19.[13] Insulin resistance pathway is also noticeable related to SARS-CoV-2 pathogenesis. Angiotensin Converting Enzyme 2 (ACE2) is a potentially important molecular link between insulin resistance and COVID-19 severity. It serves as the ligand through which coronaviruses like SARS-CoV-2 bind to their target cells. [14]

It has been shown that SARS-CoV-2 endocytosis occurs through a clathrin-mediated pathway which is regulated by the PI3K/AKT signaling. One of the possible therapeutic targets is the phosphatidylinositol 3-kinase (PI3K)/AKT signaling pathway, which is involved in various aspects of virus entry into the cell and the development of immune responses.[15] Notch signaling seems likely responsible for exacerbating COVID-19-associated coagulopathy and can be used as a possible target therapy to prevent SARS-Co-V-2 infection from progressing to cardiac diseases.[16]

In addition, we were able to identify the association of the COVID-19 pathogenesis in mismatch repair pathway, proteoglycans in cancer, NAFLD(Non-alcoholic fatty liver disease), coagulation cascades, TCA.[17][18][19][20][21]

Validation through the SMILES convolution fingerprint (SCFP) model

To further verify the pharmacological relationship between COVID-19 and MERS and influenza obtained through the previous method, we tried to find out how effective two drugs will be against COVID-19 through the SMILES convolution fingerprint (SCFP) model. We first trained and evaluated this SCFP model by using five-fold cross validation, with several indicators such as memory usage, computation time. As we mentioned above, we divided a thousand canonical SMILES data into test and evaluation dataset. We continued the training until 200 epochs while measuring the ROC-AUC for validation, the ROC-AUC of evaluation was converged at around 0.7919.

In sequence, β-D-glucopyranoside and benzoic acid that were identified the relationship with COVID-19 through previous experiments were applied to this trained SCFP model based on other chemical compounds identified through PubChem. The SCFP model predicted that 0.6775259 and 0.4953517 respectively. It means the probability of being active as a treatment is 67.75% for β-D-glucopyranoside and 49.53% for benzoic acid.

Discussion

Methods that can compensate for different experimental environments should be further studied in order to compare the RNA sequencing data of COVID-19 with the experimental data, including information before and after drug treatment for other diseases. In this study, since the confounding variables between the data were not refined, comparison of raw data of diseases was not performed, and pathway analyses were attempted by using data from previous studies. Therefore, a method was adopted to prove the reliability of the results through additional literature searches for each pathway, but problems such as high adjusted p value in the KEGG pathway analyses still remained. For more sophisticated results without in vivo experiments, it is necessary to study a calibration method that enables comparison between previous studies under different conditions.

//conclusion : β-D-glucopyranoside 가 더 높게 나온 것에 대한 부연설명

The number of Covid-19 relaxed chemical compound datasets was considerably small to sufficiently learn our SCFP model, and it did not show such high accuracy. In this case, the model can be improved slightly by fine-tuning the CNN layer or adjusting the batch size appropriately. This modification about the architecture of the model is also helpful to improve the performance. Nevertheless, one of the fundamental difficulties in fully applying pharmaceutical data to this model is that although data on-target can be obtained, there is not enough research on off-target, therefore the data quality to develop this model is inevitably not fully intact.

Even though the SMILES notation let us machinery calculation for chemical compounds, the limitations of the canonical SMILES format have also existed. The constraints are that there is no suitable standard way to generate a canonical representation and in that treatment of stereochemistry, one of the most difficult aspects of the problem is not fully explained. In this respect, it is challenging to reflect chemical properties in the process of converting chemical compounds data into a computable form.

Considering the time and initial expense required to develop any of these options, in silico drug discovery seems to be one of the efficient approaches. However, further studies should be conducted in the laboratory to validate the role of these approaches for drug discovery. Even though numerous compounds have been proposed computationally as enable treatments of COVID-19, not a few compounds and drugs have been revealed as little or no clinical supporting data or rationale. Therefore, to further improve the limitation, we performed to validate computational drug targeting by using integrated biological resources and a CNN model based on chemical representation. Likewise, the use of carefully processed prior biological data and suitable deep-learning model with accurate and well-classified data can enable the successful experimental discovery of viable drug candidates.

References

1.Brunetta, E., Folci, M., Bottazzi, B. et al. Macrophage expression and prognostic significance of the long pentraxin PTX3 in COVID-19. Nat Immunol 22, 19–24 (2021).

2.1) Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." Journal of chemical information and computer sciences 28.1 (1988): 31-36.

3.1) OpenSMILES Home Page. http://opensmiles.org. Accessed Oct. 15th 2021.

4.1) H. Shin et al., "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285-1298, May 2016, doi: 10.1109/TMI.2016.2528162.

5.1) Hirohara, M., Saito, Y., Koda, Y. et al. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinformatics 19, 526 (2018). https://doi.org/10.1186/s12859-018-2523-5

6.1) Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., & Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning. arXiv preprint arXiv:1705.08292.

7.Bossù, P., Toppi, E., Sterbini, V., & Spalletta, G. (2020). Implication of aging related chronic neuroinflammation on COVID-19 pandemic. Journal of Personalized Medicine, 10(3), 102.

8.Breikaa, R. M., & Lilly, B. (2021). The Notch Pathway: A Link Between COVID-19 Pathophysiology and Its Cardiovascular Complications. Frontiers in Cardiovascular Medicine, 8, 528.

9.Yang, M., & Lai, C. L. (2020). SARS-CoV-2 infection: can ferroptosis be a potential treatment target for multiple organ involvement?. Cell death discovery, 6(1), 1-6.

10.Vatsalya, V., Li, F., Frimodig, J. C., Gala, K. S., Srivastava, S., Kong, M., ... & McClain, C. J. (2020). Therapeutic Prospects for Th-17 cell immune storm syndrome and neurological symptoms in COVID-19: Thiamine efficacy and pharmacokinetic profile. medRxiv.

11.Al Sulaiman, K., Aljuhani, O., Al Dossari, M., Alshahrani, A., Alharbi, A., Algarni, R., ... & Al Ghamdi, G. (2021). Evaluation of thiamine as adjunctive therapy in COVID-19 critically ill patients: a two-center propensity score matched study. Critical Care, 25(1), 1-8.

12.Qin, Z., Xiang, K., Su, D. F., Sun, Y., & Liu, X. (2021). Activation of the cholinergic anti-inflammatory pathway as a novel therapeutic strategy for COVID-19. Frontiers in immunology, 11, 3870.

13.Sazgarnejad, S., Yazdanpanah, N., & Rezaei, N. (2021). Anti-inflammatory Effects of GLP-1 in Patients with COVID-19. Expert Review of Anti-infective Therapy, (just-accepted).

14.Finucane, F. M., & Davenport, C. (2020). Coronavirus and obesity: could insulin resistance mediate the severity of Covid-19 infection?. Frontiers in public health, 8, 184.

15.Khezri, M. R. (2021). PI3K/AKT signaling pathway: a possible target for adjuvant therapy in COVID-19. Human Cell, 34(2), 700-701.

16.Breikaa, R. M., & Lilly, B. (2021). The Notch Pathway: A Link Between COVID-19 Pathophysiology and Its Cardiovascular Complications. Frontiers in Cardiovascular Medicine, 8, 528.

17.Zhang, Q., Chen, C. Z., Swaroop, M., Xu, M., Wang, L., Lee, J., ... & Ye, Y. (2020). Targeting heparan sulfate proteoglycan-assisted endocytosis as a COVID-19 therapeutic option. bioRxiv.

18.Haque, F., Lillie, P., Haque, F., & Maraveyas, A. (2021). Deficient DNA mismatch repair and persistence of SARS-CoV-2 RNA shedding: a case report of hereditary nonpolyposis colorectal cancer with COVID-19 infection. BMC Infectious Diseases, 21(1), 1-4.

19.Portincasa, P., Krawczyk, M., Smyk, W., Lammert, F., & Di Ciaula, A. (2020). COVID-19 and non-alcoholic fatty liver disease: two intersecting pandemics. European journal of clinical investigation, 50(10)

20.Wang, X., Sahu, K. K., & Cerny, J. (2021). Coagulopathy, endothelial dysfunction, thrombotic microangiopathy and complement activation: potential role of complement system inhibition in COVID-19. Journal of thrombosis and thrombolysis, 51(3), 657-662.

21.Li, B. W., Fan, X., Cao, W. J., Tian, H., Wang, S. Y., Zhang, J. Y., ... & Shui, G. H. (2021). Systematic Discovery and Pathway Analyses of Metabolic Disturbance in COVID-19. Infectious Diseases & Immunity, 1(2), 74.