DCPC: Drug Candidates for the Prevention of COVID-19 Database

Ahmad Afif Supianto ‡
Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia
ahma063@brin.go.id

Rizky Nurdiansyah ‡
Department of Bioinformatics, Indonesia International Institute for Life Sciences, Indonesia
rizky.nurdiansyah@i3l.ac.id

Chia-Wei Weng ‡
Institute of Medicine, Chung Shan Medical University, Taiwan
jeff19811029@gmail.com
‡ co-first authorship

Heni Dwi Windarwati
Department of Mental Health Nursing, Faculty of Medicine, Universitas Brawijaya, Indonesia
henipsik.fk@ub.ac.id

Raden Sandra Yuwana
Research Center for Data and Information Sciences,National Research and Innovation Agency, Indonesia
rade018@brin.go.id

Andria Arisal
Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia
andria.arisal@brin.go.id

Vicky Zilvan
Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia
vick001@brin.go.id

Hilman Ferdinandus Pardede
Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia
hilm001@brin.go.id

Chien-Hung Huang
Department of Computer Science and Information Engineering, National Formosa University, Taiwan
chhuang@nfu.edu.tw

Ana Heryana
Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia
anah002@brin.go.id

Dikdik Krisnandi
Research Center for Data and Information Sciences, National Research and Innovation Agency, Indonesia
dikd001@brin.go.id

Ka-Lok Ng *
Department of Bioinformatics and Medical Engineering, Asia University, Taiwan
Corresponding author:
ppiddi@gmail.com

***Abstract*—The world immediately studied Coronavirus Disease 2019 (COVID-19) and raced towards finding the cure and developing an effective treatment. An automated approach is needed to discover drug candidates and provide those data to facilitate clinical trials in saving time and only focusing on the candidates which potentially become the cure for COVID-19. We propose the Drug Candidates for the Prevention of COVID-19 (DCPC) Database. DCPC Database provides a list of candidates of potential drugs for the prevention of COVID-19 based on disease-drug associations which are automatically discovered from biomedical literature. DCPC database is an integrative structural database, which involves a chemical database repository, such as PubChem and DrugBank to ensure that drug compound candidates have a standard representation of compounds. The database provides keyword-chosen categories and a determination of minimum supported articles for search, a list of drug candidates in the sorted table followed by the detail for each candidate, and a download feature. The keyword category consists of three keywords, they are Chinese herbal compounds, Indian medicinal plants, and Indian medicinal plants & diabetic treatment herbs. Each candidate links to an article in the biomedical literature and to a page of the compound structure visualization. DCPC is freely available at https://dcpc.brin.go.id/dcpc/.**

Keywords— SARS-CoV-2, COVID-19, text mining, disease-drug associations, herbal medicine

I. Introduction

Over the past two years, the SARS-Cov-2 coronavirus has caused many deaths. How to effectively improve our immunity to prevent or even treat this disease is an urgent and important issue. Many drugs have been proposed, but they are scattered in different studies or literature. Therefore, how effectively search for a massive of literature and integrate them into useful information is extremely critical. Since it takes a long time to develop new drugs, it will be more useful to consider approved drugs and reposition that for treating COVID-19, it is because the side effects of repositioning drugs are not that harmful.

To cope with COVID-19, several drug databases have been published [1-3]. DockCoV2 collects FDA-approved and Taiwan NHI-approved drug information and performs molecular docking with seven SARS-CoV-2 target proteins [1]. COVID19 Drug Repository uses a text-mining approach to document FDA-approved drugs plus a variety of information with detailed descriptions [2]. The Anti-SARS-CoV-2 Repurposing Drug Database records both in vivo and in vitro data to evaluate the ability of in vivo drugs to treat COVID-19 [3]. The AIM (Artificial Intelligence in Medicine) project makes use of machine learning methods to identify potential drugs for treating COVID-19 (https://covid19-help.org/), but the website did not provide further drug information after 2020. The aforementioned databases do not specifically target herbal medicines.

We perform a large-scale collection of potential drugs from the existing literature by using six categories of keywords, which is related to the use of Chinese medicinal compounds, Indian medicinal compounds, and diabetic treatment herbs with 'MeSH' and 'Field' keywords. The keywords we used for diseases are SARS, Coronavirus, HIV, MERS, and Ebola. Then, we used clustering methods to determine the drug-disease association and infer potential

DOI 10.1109/BIBE55377.2022.00034

2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE) | 978-1-6654-8487-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBE55377.2022.00034

drugs. Finally, we constructed a database, DCPC, and developed a web page to present our results.

II. Materials and Methods

A. Data Collection

This study collects abstracts from articles available in the biomedical literature. We used PubMed (https://pubmed.ncbi.nlm.nih.gov) as the data source which is one of the largest biomedical databases with an average growth of two articles per minute. The collected abstracts were then annotated using Bidirectional Encoder Representations from Transformers for Biomedical (BioBERT) text-mining [4] to identify and obtain a collection of diseases and drug names.

To ensure that the names of drug candidates are drugs with compounds that meet the standards for chemical information processing, we use the Simplified Molecular Input Line Entry System (SMILES), proposed in [5]. We utilize PubChem (https://pubchem.ncbi.nlm.nih.gov) to identify drug names based on SMILES. We collect drug candidates available on PubChem and discard drugs that are not available on PubChem. Finally, we collected a list of COVID-19 drugs from DrugBank (https://go.drugbank.com/covid-19). We used the collected drugs in DrugBank to compare with our findings. If it is available on DrugBank, we collect it as a drug candidate.

B. Database Development

The DCPC was developed and can be accessed through a web-based application that consists of three components, they are database management interface, server interface, and graphical client interface (see Figure 1). In the database management interface, it is used MySQL database management system v8.0.25. In the server interface, it is built using Django web server framework v3.2.5 and Python v3.9.12 for server-side programming. In terms of graphical client interface, it is involved HTML5, JavaScript, and CSS3 to produce an interactive web interface that can be accessed using a web browser such as Google Chrome, Mozilla Firefox, or Microsoft Edge.

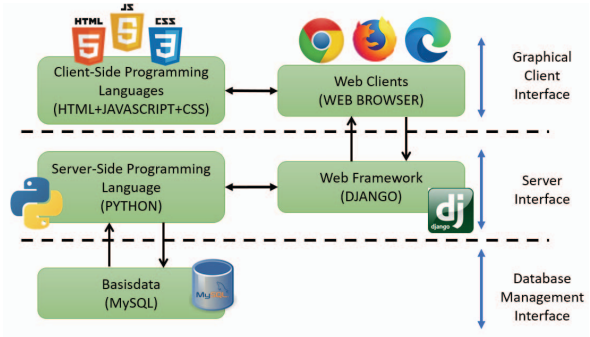

Fig. 1. Architecture of DCPC database that developed as web-based application.

The architecture of DCPC is broadly categorized into graphical client interface (front-end), server interface (back-end), and database management interface (storage). In a graphical client interface, it consists of user interactive web pages, which are accessible to the users (to view the dashboard, search for drug candidates, and download data) and administrator (who can manage data, update a collection of drug candidates from the new collection of the biomedical literature, and server monitoring). The site administrator has direct access to the storage interface and several modules in the DCPC, through an administrator login page.

The workflow diagram of the DCPC database representing the automatic disease-drug discovery, database architecture, database schema, entity-relationship model of the database, and link information to the three open datasets are shown in Figure 2. The drug-candidates discovery process is accomplished through the use of BioBERT, a text mining tool. We generate disease-drug relationship to obtain the correspondence between the disease and drug found in the abstract of the articles, and clustering the relationships based on their similarities by using Term Frequency-Inverse Document Frequency (TF-IDF) as the features and employ Agglomerative Hierarchical Clustering (AHC) as the clustering algorithm [6]. This list of the disease-drug relationships is used to extract the potential drugs. This extraction phase utilizes PubChem as a validation source for drug compounds and is continued by utilizing DrugBank as a validation source that the drug compounds are drugs that have the potential to treat COVID-19. The list of drug candidates that pass those two validations is drug candidates that pass the criteria to be analyzed in the docking calculation. Those drug candidates are then stored in the DCPC database. The attribute of "pubchem_id" is linked to the PubChem website, the attribute of "drugbank_id" is linked to the DrugBank website, and "article_ids" is linked to the PubMed website.

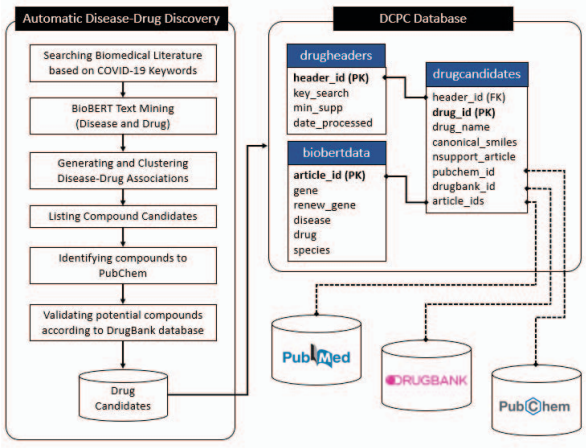

Fig. 2. Workflow diagram and database schema of the DCPC database.

C. Molecular drug docking calculation

We have selected four protein domain sites of the SARS-CoV-2 spike protein and performed drug docking calculations to rank the binding affinities of the predicted compounds. The PDB codes of the four domains are 6VW1, 6LXT, 6YB, and 6VXX. The SMILE structures of the drugs can be obtained from the PubChem database. The drug docking calculation was conducted by using the AUTODOCK VINA package. Lower binding energy indicates that the drug binds tightly to the viral proteins.

III. Result and Discussion

DCPC is an automatic disease-drug discovery and integrated database. It serves cluster-based text mining for extracting drug candidates for the prevention of COVID-19 from biomedical literature as well as provides an open dataset of the finding. Currently, the DCPC database has published

thirty records of drug candidates (with minimum support of articles equal to 3 and the lowest binding energy obtained from molecular docking), which are available to be downloaded by users. The downloaded files list the information about drug names, canonical SMILES, number of articles, and set of PubMed article IDs. The finding is obtained by the following setting parameters:

- AHC algorithm - distance = Euclidean distance.
- AHC algorithm - linkage method = average linkage.
- AHC algorithm - cutoff = 2.6.
- Minimum support of articles = 3.

Although the downloaded files are based on the minimum support of 3 articles, the website listed all drug candidates for the minimum support parameter of 3 to 12 articles.

The web interface of DCPC was designed for users without much technical skill so that they can easily access the information. For the first time, users are presented with a dashboard-like view on the "Home" page. They can select other available features in the set of menus at the top of the website, just below the header. The second feature is "About", which contains information about DCPC (background, objective, and method). The next feature is "Browse", which allows users to explore drug candidates based on two searching parameters according to their wishes, namely the keyword category and the minimum number of articles stating that the drug is worthy of being a COVID-19 drug candidate. The next would be the "Drug Candidates" feature, which is a core feature of the DCPC database where this feature contains a list of drug candidates from the results of the discovery using a text mining method. Another important feature is the "Download" feature. The "Contact" feature contains the contact persons, and the last feature is the "Admin Page" which is a page for Administrators to manage data, parameters setting, and process discovery to update the data.

The brief descriptions of the web page interfaces are elaborated in the following subsections:

1. Home page

The Home page contains statistical information from the findings with the parameters described previously. Information is presented by informative charts. First, a bar chart that contains information about the number of all drug candidates found compared to the number of drug candidates that have the potential to prevent or treat COVID-19. It displays according to the number of minimum supported articles. Secondly, the information that lists top-10 disease-drug relationships mentioned in the articles. The chart is presented in the form of a doughnut chart, displaying the name of the relationships between the disease and the drug. A screenshot of the home page interface is shown in Figure 3.

2. Browse page

The Browse page is intended for users who can freely explore drug candidates according to their desired searching criteria selection. There are two filtering levels, namely the search keyword category and minimum support. The search keyword category consists of six items, they are Chinese herbal compounds by Field keywords, Chinese herbal compounds by MeSH keywords, Indian medicinal plants by Field keywords, Indian medicinal plants by MeSH keywords, Indian medicinal plants & diabetic treatment herbs by Field keywords, and Indian medicinal plants & diabetic treatment herbs by MeSH keywords. While the minimum supported articles filter consists of the numbers 3 to 12.

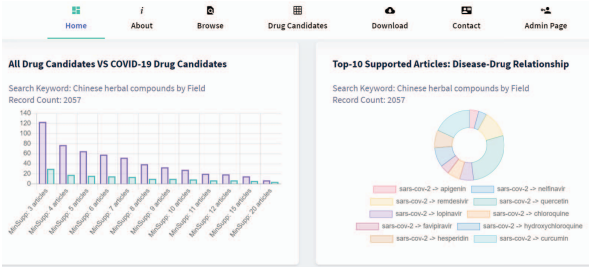

Fig. 3. Screenshot of the Home page interface.

3. Drug Candidates page

This page displays a list of drug candidates in a tabular form. The table is equipped with pagination which allows users to view a list of 5 drugs per page (see Figure 4). The table contains information about Drug Name, Canonical Smiles, Support, PubChemID, DrugBankID, Mean Binding Energy (KJ/mol), and Actions. In the Actions column, there is a button that when clicked will display a detailed information page of the selected drug. The information refers to the hyperlink to the PubMed articles and the results of BioBERT mining, which include PubMed ID, Gene, Renew Gene, Diseases, Drugs, Species, and the Article's abstract. In the abstract, annotations are displayed for all diseases and drugs contained in the abstract. This detail page can be seen in Figure 5.

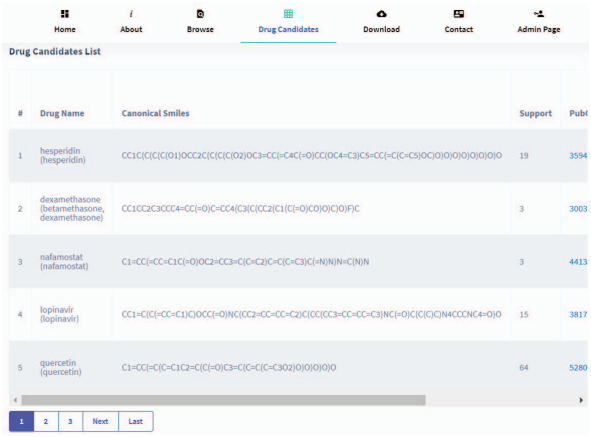

Fig. 4. Screenshot of the Drug Candidates page interface.

4. Download page

The Download page allows users to download the drug candidate lists, which consist of six download options according to the aforementioned six different searching keyword categories. Users are allowed to download each category in the Excel file format (.xlsx). The file consists of

records with the attributes of drug_name, canonical_smiles, support, and pubmed_ids.

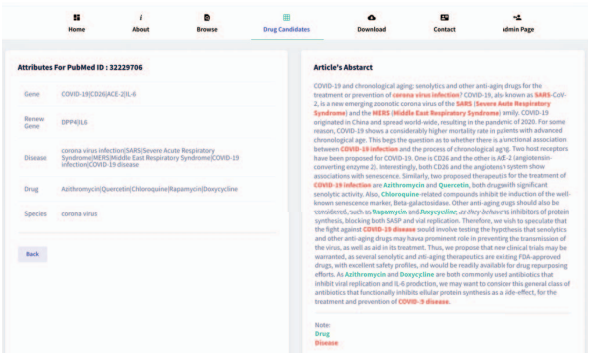

Fig. 5. Screenshot of the information of the Drug Candidates' detail page.

5. Admin Page

The Admin pages are used to manage data, adjust the data presented to users, determine setting parameters, and update drug candidates data that can be downloaded by users. Prior to entering the Administrator area, the webpage interface is preceded by a Login page and only those who have access as Administrator are allowed to access this Admin page.

IV. Conclusion

DCPC Database is an integrated, comprehensive, and open access resource for providing potential drugs to prevent and/or treat Coronavirus diseases. All entities in the DCPC are integrated by the related entries to provide more information. DCPC strives to provide a list of drug candidates that can be easily accessed and downloaded. DCPC database is still under development, as we are currently focused on increasing data accuracy by discovering disease-drug associations based on an automatic approach as well as it will not be limited to the PubMed dataset to provide more information. In future study, we would like to identify and analyze the related biological processes or pathways of the drug-target proteins. We are constantly striving to complete the process and provide an online server.

Acknowledgment

The works are supported by Research Program (DIPA Rumah Program) at the Research Organization for Life Sciences and Environment, the National Research and Innovation Agency (BRIN), Indonesia (grant number: 9/III/HK/2022), Dr. Ka-Lok Ng work is supported by National Science and Technology Council (NSTC), Taiwan (grant number: NSCT 111-2221-E-468-014) and the Asia University (grant number: ASIA-110-CMUH-12).

References

[1] T. F. Chen et al., "DockCoV2: a drug database against SARS-CoV-2," (in eng), Nucleic Acids Res, vol. 49, no. D1, pp. D1152-d1159, Jan 8 2021, doi: 10.1093/nar/gkaa861.

[2] D. Tworowski et al., "COVID19 Drug Repository: text-mining the literature in search of putative COVID19 therapeutics," Nucleic Acids Research, vol. 49, no. D1, pp. D1113-D1121, 2020, doi: 10.1093/nar/gkaa969.

[3] X. Zhang et al., "Anti-SARS-CoV-2 Repurposing Drug Database: Clinical Pharmacology Considerations," (in eng), CPT Pharmacometrics Syst Pharmacol, vol. 10, no. 9, pp. 973-982, Sep 2021, doi: 10.1002/psp4.12681.

[4] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics," 36(4), 1234-1240.

[5] Weininger D. "SMILES, a chemical language and information system: 1. Introduction to methodology and encoding rules." Journal of Chemical Information and Computer Sciences. 1988;28(1):31-6.

[6] Bouguettaya, A., et al., "Efficient agglomerative hierarchical clustering. Expert Systems with Applications," 2015. 42: p. 2785-2797.