D3Similarity: A ligand-based approach for predicting drug targets and for virtual screening of active compounds against COVID-19

Zhengdan Zhu[1,2,#], Xiaoyu Wang[1,3,#], Yanqing Yang[1,2,#], Xinben Zhang[1,#], Kaijie Mu[1,4], Yulong Shi[1,2], Cheng Peng[1,2], Zhijian Xu[1,2,*], Weiliang Zhu[1,2,*]

[1]CAS Key Laboratory of Receptor Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China.

[2]School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

[3]College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China.

[4]Nano Science and Technology Institute, University of Science and Technology of China, Suzhou, Jiangsu, 215123, China.

[#]These authors contributed equally to this work.

[*]To whom correspondence should be addressed.

Phone: +86-21-50806600-1201 (Z.X.), +86-21-50805020 (W.Z.),

E-mail: zjxu@simm.ac.cn (Z.X.), wlzhu@simm.ac.cn (W.Z.).

ORCID:

Zhijian Xu: 0000-0002-3063-8473

Weiliang Zhu: 0000-0001-6699-5299

Abstract

Discovering efficient drugs and identifying target proteins are still an unmet but urgent need for curing COVID-19. Protein structure based docking is a widely applied approach for discovering active compounds against drug targets and for predicting potential targets of active compounds. However, this approach has its inherent deficiency caused by, e.g., various different conformations with largely varied binding pockets adopted by proteins, or the lack of true target proteins in the database. This deficiency may result in false negative results. As a complementary approach to the protein structure based platform for COVID-19, termed as D3Docking in our recent work, we developed the ligand-based method, named D3Similarity, which is based on the molecular similarity evaluation between the submitted molecule(s) and those in an active compound database. The database is constituted by all the reported bioactive molecules against the coronaviruses SARS, MERS and SARS-CoV-2, some of which have target or mechanism information but some don't. Based on the two-dimensional and three-dimensional similarity evaluation of molecular structures, virtual screening and target prediction could be performed according to similarity ranking results. With two examples, we demonstrated the reliability and efficiency of D3Similarity for drug discovery and target prediction against COVID-19. D3Similarity is available free of charge at https://www.d3pharma.com/D3Targets-2019-nCoV/D3Similarity/index.php.

1. Introduction

The novel coronavirus pneumonia (COVID-19) induced by SARS-CoV-2 (previously named as 2019-nCoV) infection[1, 2] has caused more than 3800 deaths as of 9 March 2020. However, there is still no effective drug approved for clinical treatment, leaving the clinic needs unmet.

Quite a number of compounds including natural products have been reported to be active against various coronavirus at different levels. For example, glycyrrhizin was found to have bioactivity in inhibiting the replication, absorption and penetration of SARS-CoV.[3] Tanshinones, which are a series of natural products derived from Salvia miltiorrhiza, were reported as inhibitors against the 3C-like, papain-like and viral cysteine proteases of SARS-Cov,[4] with one of the compounds in this series exhibit nanomolar level activity ($IC_{50} = 0.8 \pm 0.2$ μM) against the papain-like protease of SARS-CoV. Several FDA approved drugs, including chloroquine, chlorpromazine, loperamide and lopinavir,[5] showed in vitro activity in the inhibition of MERS-CoV replication under low-micromolar level (3-8 μM). For the novel coronavirus SARS-CoV-2, compounds represented by remdesivir and chloroquine[6, 7] have also demonstrated promising in vitro or even potential in vivo bioactivity. Although the active mechanism of some of the reported bioactive molecules have been explored, there may still be a large proportion of compounds, of which the corresponding target protein and the mechanism behind the bioactivity remain to be revealed.

Previously, we embedded a structure-based module named D3Docking in the D3Targets-2019-nCoV web server,[8, 9] which utilized molecular docking to explore the potential protein-ligand binding energies, and has already been visited for more than 2200 times by worldwide researchers. Though we tried to consider the influence of conformations and pockets in the D3Docking module, it is very difficult to comprehensively and precisely predict all the possible conformations and druggable pockets in a structural-based approach using molecular docking, and thus may lead to false negative prediction. Additionally, some of the compounds may interact with the target proteins not included in the database of D3Docking. Therefore, developing another parallel approach for target identification and virtual screening is necessary as an alternative to the structure-based scheme.

Here we presented a ligand-based approach, named D3Similarity, to predict active compounds against either SARS-CoV-2 or COVID-19 (considering the involved human proteins), and to identify

the potential target proteins for molecules with potential bioactivity in a scheme that is irrelevant to the reliability and accuracy of protein 3D structures. This was realized by evaluating the molecular similarity between the input molecules and the active compounds in the D3Similarity database. We hope D3Similarity would provide another efficient way for target identification and virtual screening to meet the need for curing COVID-19.

2. Materials and methods

2.1 Preparation of the ligand-based database. A total of 157 molecules with potential bioactivity in the treatment of different types of coronavirus infection were collected to construct the ligand-based database, involving targets of both viral (including SARS, MERS, SARS-CoV-2) and human proteins. Among the ligands of the database, 138 molecules were summarized in literatures[10, 11] and were reported to have bioactivity against coronavirus infection. The remaining 19 molecules were recorded in the ChEMBL[12] database, which were indexed by the Uniprot ID of the associated proteins for the reported 138 molecules, and were thus deduced to have anti-coronavirus-infection bioactivity. The ligand-based database of D3Similarity will be continuously updated in our future work.

2.2 Preprocessing of small molecules. All the small molecule files, including that inputted by the user and those already existing in the ligand-based database, would be preprocessed under the identical workflow. Generally, the small molecule file would be first transformed to the mol format with Open Babel;[13] following optimization under the MMFF94 force field with the RDKit package;[14] the mol file outputted by RDKit with optimized structures would then be transformed to the mol2 format again with Open Babel to be prepared for the molecular similarity evaluation task.

Notably, we found that small molecules involving the nitro group (-NO_2) could not be well handled by RDKit if essential information besides atomic coordinates and bonds is missing in the structure file. This essential information involves additional atomic charge definition of the nitro group, termed as the “unity atom attributes” in a qualified mol2 format file, which put one positive charge unit on the nitrogen atom while one negative charge unit on one of the oxygen atoms. Thus, we recommend that molecules containing nitro groups to be preprocessed with a SMILES based workflow. The input molecule will first be transformed to the SMILES string if our program determines that the molecule

is substituted with the nitro group. Subsequently, the SMILES of this small molecule would be revised (to include the "unity atom attributes" of the nitro group) and used as the input, following optimization using RDKit with the MMFF94 force field to get the three-dimensional structure, and finally be transformed to the mol2 format with Open Babel.

2.3 Evaluating the 2D molecular similarity. The 2D molecular similarity were evaluated based on the Tanimoto coefficient (Tc) values between the SMILES of the input structure and the sdf file containing all molecules in the database, which was obtained using Open Babel based on the mol2 files generated in part **2.2**. The Tc values were calculated with Open Babel using the default FP2 fingerprint.

2.4 Evaluating the 3D molecular similarity. The 3D molecular similarly between the input molecule and those in the ligand-based database was evaluate based on the mol2 files generated in part **2.2** using MolShaCS,[15] which is a computational tool to assess the molecular shape and charge similarity between two molecules. Parameters used in the evaluation task (Table 1) were set to the recommended values as mentioned in the MolShaCS manual.

Table 1. Parameters used in the molecular similarity evaluation task in MolShaCS.

Parameter name	Value
minimizer	nlopt_mma
align_molecules	yes
timeout	60
write_coordinates	yes
mol2_aa	no
box_size	30.0
step	1.0E-5
tol	1.0E-4
delta	1.0E-5

2.5 Ligand-based virtual screening. The ligand-based virtual screening would be conducted based either on target protein-related compounds or on the active compounds without any target information. Molecular similarity would be evaluated between the molecules in the input sdf or mol2 file and those

in a subset of the ligand-based database. The output result would simply be ranked by 2D and 3D molecular similarity for all involved pairs of input molecule and database ligands, and thus offer suggestions in choosing promising molecules for further experimental exploration.

3. Results and discussion

3.1 Overview of molecules and potential target proteins included in the database. More than 30 targets were involved for the 157 potential bioactive molecules that are contained in our ligand-based database. Inhibitors for the 3C-like and papain-like proteases account for the two largest proportions among all involved molecules (Figure 1). Notably, molecules with multiple targets were also counted for multiple times in the pie chart plotted in Figure 1. Details of ligand structures and the associated information for the target(s) are provided on the web page (see the CoViLigands module, https://www.d3pharma.com/D3Targets-2019-nCoV/CoViLigands/2019-nCoV.php).

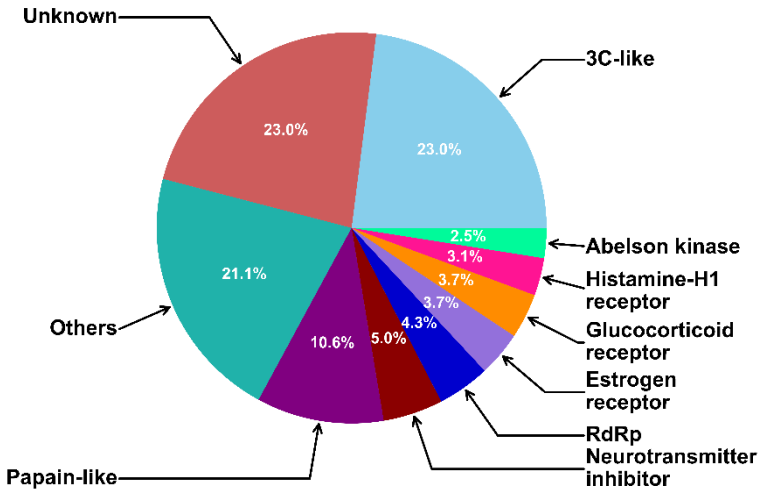

Figure 1. Pie chart for the percentage of associated targets or types for small molecules composing the ligand-based database.

3.2 Input and output. D3Similarity is provided free of charge for registered users of the D3Targets-2019-nCoV web server (https://www.d3pharma.com/D3Targets-2019-nCoV/D3Similarity/index.php). A graphical interface of the target identification module is shown in Figure 2. We recommend that the users submit the input structure in common file formats such as mol2 and sdf to ensure the input file could be well handled by D3Similarity. Usually the evaluation of molecular similarity between the

submitted molecule and those in the database would last for several minutes after the beginning of the calculation before the output result is returned, in which the information of the molecular structures and associated target(s) for top-ranking ligands will be provided on the web page.

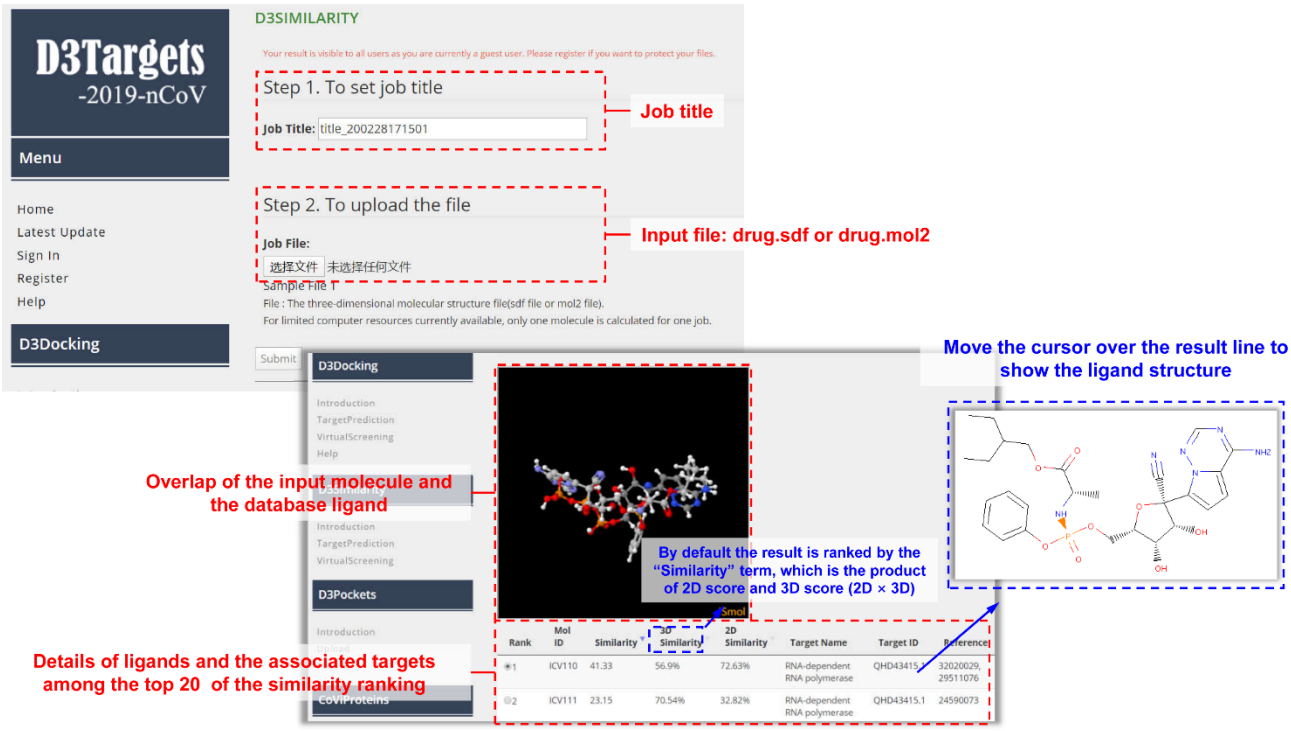

Figure 2. Graphical interface for input and output of the target identification module of D3Similarity.

A similar graphical interface is also provided for the virtual screening module of D3Similarity. In this module, the users must provide the input database file in sdf or mol2 format to guarantee the program could correctly split different molecules out of the input database. An error would occur if provided with a database file in other formats. Usually the calculation would last for several minutes for each molecule in the input database, and thus, the total running time would depend on the size of the input database. In the output result, molecules in the input database would be labeled as "MOL_1", "MOL_2"… based on the order in which they appear in the input file. By default, the results would be ranked by molecular similarity between the input compound and the database ligand. However, in our future update, we would consider to involve the function to rank the result based on bioactivity of involved ligands in the database.

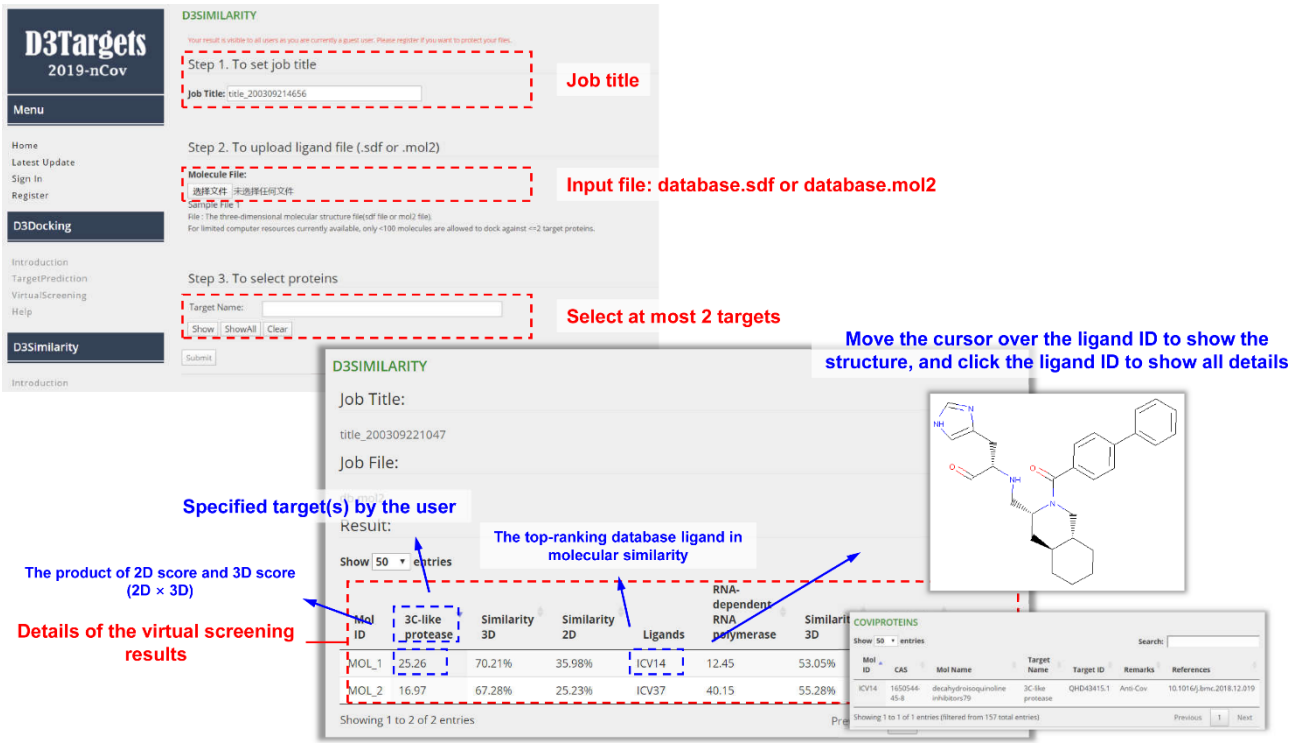

Figure 3. Graphical interface for input and output of the virtual screening module of D3Similarity.

3.3 Evaluation of different ranking methods. As mentioned above, both 2D and 3D similarity evaluation were conducted in our ligand-based module. However, considering that the SMILES-based 2D scheme may lose sight of the molecular geometry while the 3D scheme would be slightly affected by the difference in molecular conformations, here, we additionally included the evaluation results ranked by the product of 2D similarity score and 3D similarity score (2D × 3D). Case studies were presented to explore the efficiency of the three ranking methods, and thus, that of D3Similarity in identifying potential targets.

Inhibitors of the 3C-like protease and papain-like protease were selected as two typical examples considering these two subsets account for the largest proportions of the ligand-based database. Molecules in the rest of the database excluding inhibitors of 3C-like protease or papain-like protease were also selected as the reference subsets to consider the potential influence of database components. Molecular similarity evaluations were then conducted between the input ligands in the subsets and in the database excluding the input ligand itself.

As shown in the pie charts of Figure 4, in the case study of 3C-like protease inhibitors, we presented the average percentage composition of 3C-like protease, unknown and other targets that correspond to

the molecules in the top 10 similarity rankings (hereinafter referred to as "top 10" targets and "top 10" molecules) using 3C-like protease inhibitors in the database as input structures. As demonstrated in the pie charts, in the evaluation results ranked by 2D similarity, 3D similarity and the 2D × 3D scheme, 3C-like protease accounts for a significantly larger percentage among the "top 10" targets for 3C-like protease inhibitors (Figure 4a-4c) than that for molecules in the reference subset (Figure 4d-4f). This observation suggested that the large proportion of 3C-like protease in the "top 10" targets for 3C-like protease inhibitors results not only from the database component itself, but also from the successful prediction of our ligand-based approach.

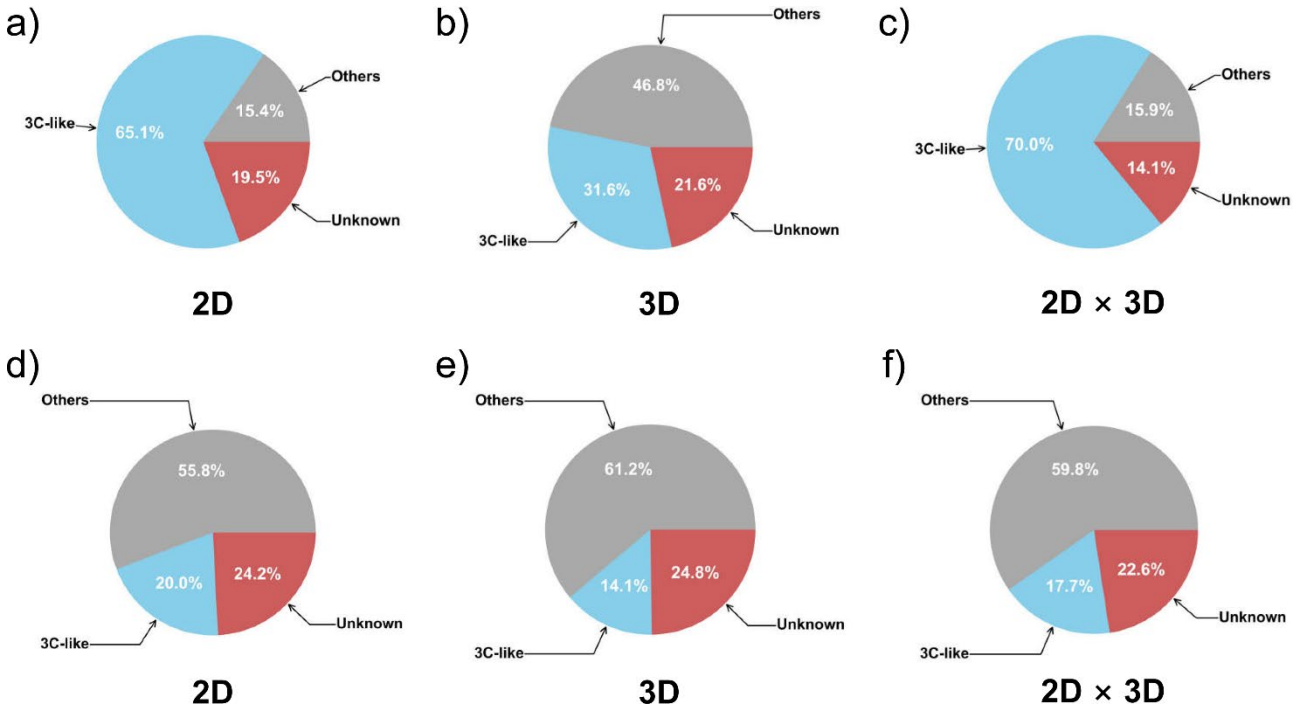

Figure 4. Case study of the 3C-like protease inhibitors using D3Similarity. Plotted pie charts are for average percentage composition of 3C-like protease, unknown and other targets that correspond to the molecules in the top 10 similarity rankings using 3C-like protease inhibitors in the database as input structures ranked by (a) 2D similarity, (b) 3D similarity, (c) the product of 2D similarity score and 3D similarity score; and using molecules in the rest of the database as input structures ranked by (d) 2D similarity, (e) 3D similarity, (f) the product of 2D similarity score and 3D similarity score.

Similar observations were also demonstrated in the case study of papain-like protease inhibitors (Figure 5). In the "top 10" targets of the similarity evaluation result, papain-like protease also accounts for a larger proportion for its reported inhibitors (Figure 5a-5c) compared with that for other molecules

(Figure 5d-5f). What's more, in general, in both two case studies, the usage of the "2D × 3D" scheme to rank the molecular similarity yielded better results than using either 2D similarity score or 3D similarity score alone, suggesting that the 2D and 3D score may complement each other after the multiplication. Thus, we recommend the users to employ the "2D × 3D" scheme as the default ranking scheme for molecular similarity.

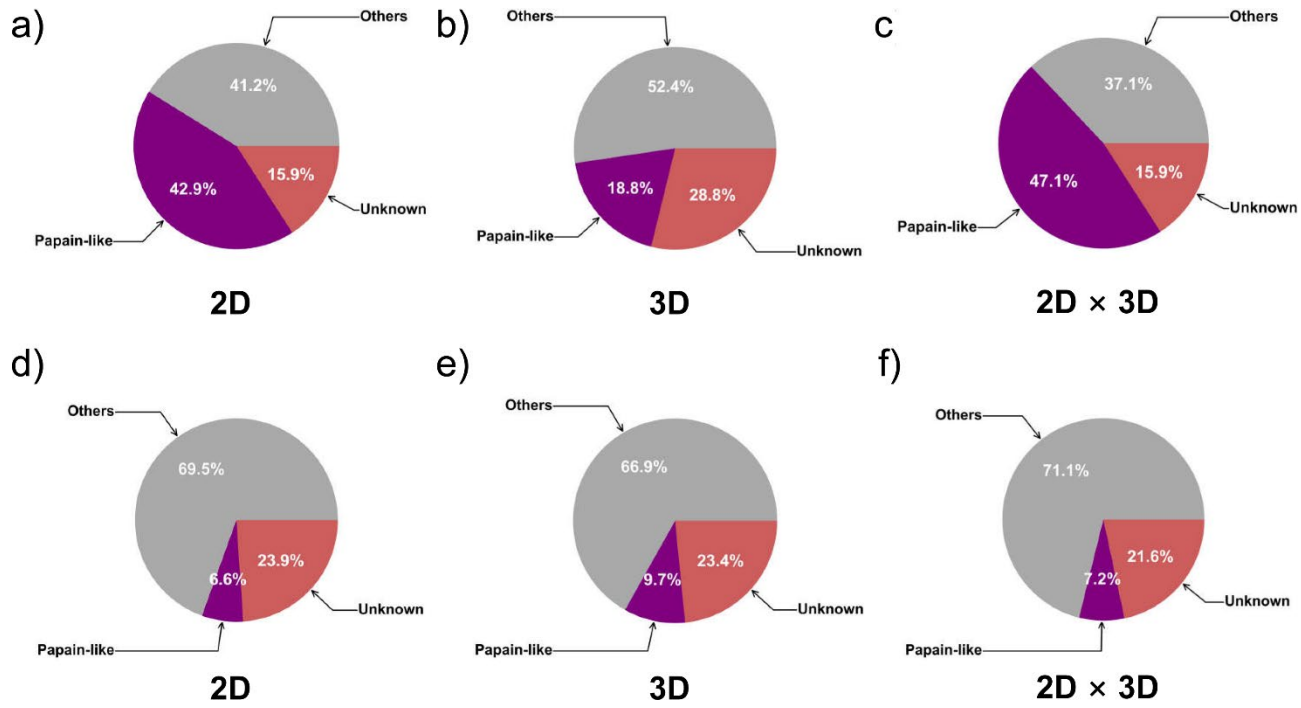

Figure 5. Case study of the papain-like protease inhibitors using D3Similarity. Plotted pie charts are for average percentage composition of papain-like protease, unknown and other targets that correspond to the molecules in the top 10 similarity rankings using papain-like protease inhibitors in the database as input structures ranked by (a) 2D similarity, (b) 3D similarity, (c) the product of 2D similarity score and 3D similarity score; and using molecules in the rest of the database as input structures ranked by (d) 2D similarity, (e) 3D similarity, (f) the product of 2D similarity score and 3D similarity score.

Overall, we believe that D3Similarity should be a complementary approach to docking based methods for ligand-based target prediction and virtual screening.

4. Conclusions

The SARS-CoV-2 infection has led to more than 3800 deaths and affected more than 90 countries

worldwide as of 9 March 2020, while no approved drug is available for clinical treatment. Virtual screening is a highly efficient approach to find potential antivirals, while identifying the potential targets is of great importance for understanding the bioactivity mechanism of both now-existing and to-be-developed molecules against the coronavirus infection. On the basis of the previously reported D3Targets-2019-nCoV web server, which has already been embedded with a structure-based module named D3Docking; in this work, we released the ligand-based module, termed as D3Similarity, which utilizes the molecular similarity evaluation with bioactive molecules with known targets or/and well-explored mechanism. 157 molecules were included in the ligand-based database, including 19 molecules indexed from the ChEMBL database, which were deduced to be bioactive. In the evaluation of different ranking methods, when applying the product of 2D similarity score and 3D similarity score (2D × 3D) to rank the results, D3Similarity correctly predicted the target of the inhibitors of 3C-like and papain proteases and outperformed the ranking results using either 2D similarity score or 3D similarity score alone. These observations demonstrated D3Similarity should be a complementary approach to docking based methods for virtual screening and target identification of potential coronavirus antivirals. We hope this ligand-based module would be helpful to the drug development against SARS-CoV-2 and other coronaviruses. The module is available free of charge for registered users of D3Targets-2019-nCoV at https://www.d3pharma.com/D3Targets-2019-nCoV/D3Similarity/index.php.

Acknowledgements

This work was supported by the National Key R&D Program of China (2017YFB0202601 & 2016YFA0502301).

References

1. Wu, F.; Zhao, S.; Yu, B.; Chen, Y. M.; Wang, W.; Song, Z. G.; Hu, Y.; Tao, Z. W.; Tian, J. H.; Pei, Y. Y.; Yuan, M. L.; Zhang, Y. L.; Dai, F. H.; Liu, Y.; Wang, Q. M.; Zheng, J. J.; Xu, L.; Holmes, E. C.; Zhang, Y. Z., A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, DOI: 10.1038/s41586-020-2008-3.

2. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi,

W.; Lu, R.; Niu, P.; Zhan, F.; Ma, X.; Wang, D.; Xu, W.; Wu, G.; Gao, G. F.; Tan, W.; China Novel Coronavirus, I.; Research, T., A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England Journal of Medicine* **2020,** *382* (8), 727-733.

3. Cinatl, J.; Morgenstern, B.; Bauer, G.; Chandra, P.; Rabenau, H.; Doerr, H. W., Glycyrrhizin, an active component of liquorice roots, and replication of SARS-associated coronavirus. *Lancet* **2003,** *361* (9374), 2045-2046.

4. Park, J. Y.; Kim, J. H.; Kim, Y. M.; Jeong, H. J.; Kim, D. W.; Park, K. H.; Kwon, H. J.; Park, S. J.; Lee, W. S.; Ryu, Y. B., Tanshinones as selective and slow-binding inhibitors for SARS-CoV cysteine proteases. *Bioorganic & Medicinal Chemistry* **2012,** *20* (19), 5928-5935.

5. de Wilde, A. H.; Jochmans, D.; Posthuma, C. C.; Zevenhoven-Dobbe, J. C.; van Nieuwkoop, S.; Bestebroer, T. M.; van den Hoogen, B. G.; Neyts, J.; Snijder, E. J., Screening of an FDA-approved compound library identifies four small-molecule inhibitors of Middle East respiratory syndrome coronavirus replication in cell culture. *Antimicrobial Agents and Chemotherapy* **2014,** *58* (8), 4875-4884.

6. Holshue, M. L.; DeBolt, C.; Lindquist, S.; Lofy, K. H.; Wiesman, J.; Bruce, H.; Spitters, C.; Ericson, K.; Wilkerson, S.; Tural, A.; Diaz, G.; Cohn, A.; Fox, L.; Patel, A.; Gerber, S. I.; Kim, L.; Tong, S.; Lu, X.; Lindstrom, S.; Pallansch, M. A.; Weldon, W. C.; Biggs, H. M.; Uyeki, T. M.; Pillai, S. K.; Washington State -nCo, V. C. I. T., First Case of 2019 Novel Coronavirus in the United States. *The New England Journal of Medicine* **2020**, DOI: 10.1056/NEJMoa2001191.

7. Wang, M.; Cao, R.; Zhang, L.; Yang, X.; Liu, J.; Xu, M.; Shi, Z.; Hu, Z.; Zhong, W.; Xiao, G., Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Research* **2020**, DOI: 10.1038/s41422-020-0282-0.

8. Chen, Z.; Zhang, X.; Peng, C.; Wang, J.; Xu, Z.; Chen, K.; Shi, J.; Zhu, W., D3Pockets: A Method and Web Server for Systematic Analysis of Protein Pocket Dynamics. *Journal of Chemical Information and Modeling* **2019,** *59* (8), 3353-3358.

9. Yulong, S.; Xinben, Z.; Kaijie, M.; Cheng, P.; Zhengdan, Z.; Xiaoyu, W.; Yanqing, Y.; Zhijian, X.; Weiliang, Z., D3Targets-2019-nCoV: A Web Server to Identify Potential Targets for Antivirals Against 2019-nCoV. *ChemRxiv*, **2020**, DOI: 10.26434/chemrxiv.11831163.v1.

10. Li, G.; De Clercq, E., Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nature*

Reviews Drug Discovery **2020**, DOI: 10.1038/d41573-020-00016-0.

11. Pillaiyar, T.; Meenakshisundaram, S.; Manickam, M., Recent discovery and development of inhibitors targeting coronaviruses. *Drug Discovery Today* **2020**, DOI: 10.1016/j.drudis.2020.01.015.

12. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2012,** *40*, 1100-1107.

13. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011,** *3*, 33.

14. Landrum, G., RDKit: Open-source cheminformatics. **2019**.

15. Vaz de Lima, L. A.; Nascimento, A. S., MolShaCS: a free and open source tool for ligand similarity identification based on Gaussian descriptors. *European Journal of Medicinal Chemistry* **2013,** *59*, 296-303.