BIOINFORMATICS AND BIOMEDICAL RESEARCH JOURNAL
2021, Vol. 4, No. 1, 1– 7
http://dx.doi.org/10.11594/bbrj.04.01.01

Research Article

Predicition of Drug-Target Interaction Using Random Forest in Coronavirus 2019 Case

Aulia Fadli[1], Annisa[2], and Wisnu Ananta Kusuma[3*]

[1,2,3]Department of Computer Science, IPB University, 16680 West Java, Indonesia
[2,3]Biopharmaca Research Center, IPB University, 16680 West Java, Indonesia

Article history:
Submission June 2021
Revised June 2021
Accepted June 2021

**Corresponding author:*
E-mail: ananta@apps.ipb.ac.id

ABSTRACT

Coronavirus disease 2019 is an infectious disease that causes severe respiratory, digestive, and systemic infections that caused a pandemic in 2019. One of the focuses of the drug development process to fight the coronavirus disease 2019 is by carrying out drug repurposing. This study was the random forest with a feature-based chemogenomics approach on the drug-target interaction data of coronavirus disease 2019. The feature extraction process was conducted on compounds and proteins using PubChem fingerprint and amino acid composition. Feature selection was performed by using XGBoost to reduce the data dimension. The random undersampling process was also carried out to solve the problem of imbalanced data in the dataset. Using the cross-validation process, The random forest model implemented in this study had a good performance in predicting DTI with an average accuracy value of 0.98, recall value of 0.92, precision value of 0.95, AUROC value of 0.95, and F1 score of 0.93. The random forest model also produced an accuracy value of 0.99, recall value of 0.93, the precision value of 0.94, AUROC value of 0.99, and F-measure of 0.94 when used to predict the original dataset (dataset without random undersampling process).

Keywords: *Coronavirus disease 2019, drug-target interaction, random forest, random undersampling, XGBoost.*

Background

Coronavirus (Cov) is a virus that causes severe respiratory, digestive, and systemic infections in humans and animals [1]. This virus has caused several major outbreaks since the early 21st century with the most recent case being the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) pandemic in 2019 [2]. On 11 February 2020, WHO designated SARS-CoV-2 as COVID-19 and defined it as a pandemic on 11 March 2020. Based on the data at the Worldometer site, as of December 12, 2020, COVID-19 cases in Indonesia had reached 605,243 cases with the number of deaths reaching 18,511 people.

COVID-19 pandemic is prompting the world to search for potential drugs to fight this virus. One of the focuses of a drug development process to fight the coronavirus disease 2019 is

How to cite:
Fadli, A., Annisa & Kusuma, W. A. (2021). Prediction of Drug-Target Interaction Using Random Forest in Coronavirus Disease 2019 Case. *Bioinformatics and Biomedical Research Journal*. *4*(1), 1 – 7. doi: 10.11594/bbrj.04.01.01

by carrying out drug repurposing [3]. Drug repurposing is a process to identify new efficacy for existing drugs and is considered an efficient and economical approach to drug development [4]. Research in drug repurposing can be done by observing the interactions of drug compounds with protein related to a disease (Drug-Target Interaction or DTI), then build a model to predict the new drug-target interactions with unknown interactions previously [5].

There are two common problems for DTI prediction: the curse of dimensionality and imbalance class 6. The curse of dimensionality problem occurred because the dataset used in DTI research generally has large dimensions. Meanwhile imbalance class occurred because the number of positive interactions on DTI is much smaller than the number of negative or unknown interactions. Imbalanced class problems can lead to the inaccurate classification of model 7. Therefore, additional handling of these problems is required. Feature selection can be used to try to solve the curse of dimensionality meanwhile random undersampling process can be used to try to solve imbalance class problem [6].

There have been researches regarding DTI in COVID-19 case before. Sulistiawan et al., the research tried to predict DTI in COVID-19 case using deep semi-supervised learning (DSSL) and deep neural network (DNN) model with chemogenomics feature-based approach on potential anti-coronaviral treatment dataset [7]. DTI search is done by first performing feature extraction on compounds and proteins. Feature extraction on compounds done using PubChem fingerprint. Meanwhile, the feature extraction process on the protein was done using dipeptide composition (DC), autocorrelation descriptors (ACD), and position-specific scoring matrix (PSSM). This research identified that the use of fingerprint for compound features and DC for protein features gave the best results for predicting DTI with an accuracy of 94%, AUROC of 0.97, and F-measure of 0.82.

In this research, a DTI prediction model will be built using the random forest algorithm. The random forest model is chosen because of its good performance in DTI prediction task [8]. The aim of this study was to implement the random forest model and evaluate its performance in predicting DTI in COVID-19 cases. Feature extraction process on compounds and protein will be done using PubChem fingerprint and amino acid composition respectively. A feature selection process is also carried out using the XGBoost algorithm to reduce data dimensions by selecting features that have no information gain value in the model. Then the random undersampling process is done to try to solve the problem of imbalanced data in the dataset. A random forest model is built to perform the learning of the dataset. Model performance will be evaluated using metrics such as accuracy, recall, precision, f-measure, and area under the ROC curve (AUROC). Another prediction model was also built using LGBM, GBM, and XGBoost models to compare the results of random forest performance.

Methods

Data

The data used in this study were compound-protein interaction data obtained from the SuperTarget and DrugBank site. The feature extraction process on compounds is done using PubChem fingerprint which resulted in 881 attributes on the compound's data. In chemoinformatics, a fingerprint is a way to represent the chemical structure of a compound [9]. The feature extraction process on protein is done using amino acid composition (AAC) which contains the percentage of amino acids in the protein sequence used in the protein data. The results of feature extraction on the protein yielded 20 attributes. The total attributes in the dataset were 904. Compounds and protein that are known to have interactions will be labeled with 1, else if pairs of compounds and proteins are not known to have interactions will be labeled with 0. The dataset has an imbalance class problem where row data labeled as 1 have 4090 instances and row data that are labeled as 0 have 138398 instances. Data examples can be seen in Table 1.

Table 1. Data examples

fc1	fc2	...	fc881	A	C	...	Y	Class
1	0	...	0	0.049	0.015	...	0.03	1
1	1	...	0	0.049	0.015	...	0.03	0
1	1	...	1	0.049	0.015	...	0.03	0
1	1	...	0	0.068	0.023	...	0.03	1
1	0	...	0	0.068	0.023	...	0.03	0

Data Acquisition

The protein data used in this study were the protein associated with COVID-19 in the human body. Information on COVID-19 protein data were obtained from literature studies and explored on the OMIM and Uniprot sites. Drug-Target Interaction Data were obtained from SuperTarget and DrugBank sites. The protein data were used as input at the site to obtain compound and protein interaction data. SuperTarget site can be accessed at http://insilico.charite.de/supertarget/ while the DrugBank site can be accessed at https://go.drugbank.com.

Data Preprocess

After collecting compound-protein interaction data on SuperTarget and DrugBank, the data were merged and redundant data were deleted. After merging and deleting redundant data, the next step was the feature extraction process on compounds and protein data. Feature extraction process on compounds and protein was conducted using PubChem fingerprint and amino acid composition respectively. In PubChem fingerprint, there is 881 chemical substructures with each substructure assigned to a specific location. The position of the corresponding substructure on the fingerprint vector is 1, otherwise, the position of the substructure is 0 [10]. The amino acid composition is the percentage of each amino acid type within a protein and can be used to explain protein information [11]. Figure 1 is presented the illustration of the feature extraction process on compounds using PubChem fingerprint [9]. Meanwhile, the feature extraction process on proteins using amino acid composition can be seen in Figure 2.

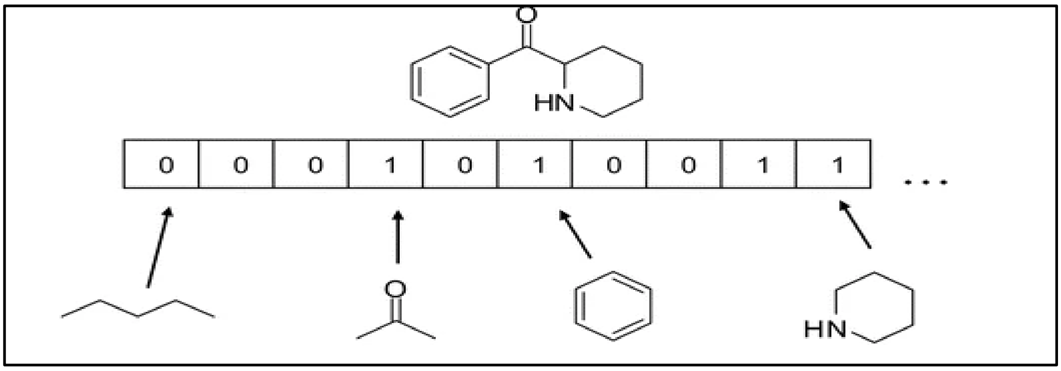

Figure 1. Illustration of fingerprint to represent the chemical structure of a compound [9]

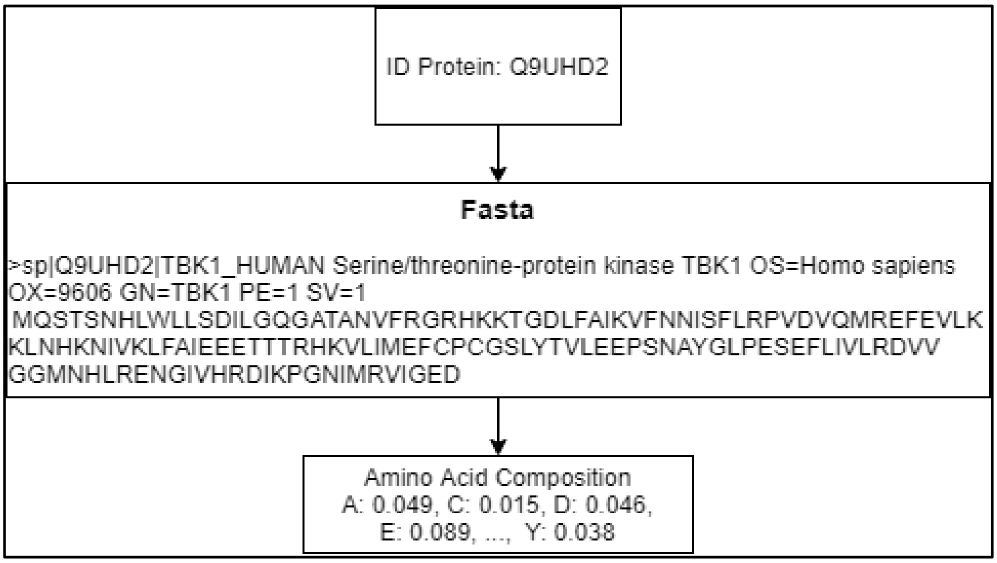
ID Protein: Q9UHD2
Fasta
>sp|Q9UHD2|TBK1_HUMAN Serine/threonine-protein kinase TBK1 OS=Homo sapiens OX=9606 GN=TBK1 PE=1 SV=1
MQSTSNHLWLLSDILGQGATANVFRGRHKKTGDLFAIKVFNNISFLRPVDVQMREFEVLK
KLNHKNIVKLFAIEEETTTRHKVLIMEFCPCGSLYTVLEEPSNAYGLPESEFLIVLRDVV
GGMNHLRENGIVHRDIKPGNIMRVIGED
Amino Acid Composition
A: 0.049, C: 0.015, D: 0.046,
E: 0.089, ..., Y: 0.038

Figure 2. Feature extraction process on proteins using amino acid composition

This dataset contains pairs of compounds and proteins that have interaction. In the model building process to predict the interaction between compounds and proteins, the model requires information on pairs of protein compounds that are known to have interactions and those whose interactions are unknown [5]. Therefore, data transformation is done by pairing all interaction pairs of unique protein IDs and compound IDs. The compound and protein feature descriptors are then combined into one input vector. Next, label 1 is given to pairs of compounds and protein that are known to have interactions, and label 0 to pairs of compounds and protein whose interactions are unknown.

DTI Prediction Model and Evaluation

DTI prediction model was built using a random forest algorithm. The feature selection process is done on the dataset using XGBoost algorithm. Feature selection aims to reduce the number of dimensions in the dataset, therefore, accelerating the computation process during the model training process. After that, a random undersampling is carried out to try to solve the problem of imbalanced data in the dataset. The random forest model is then trained using an undersampling dataset. At the model training stage, the parameter tuning process is carried out to find the optimal parameters with GridSearchCV. A parameter to be explored including max_depth (the depth of the tree), max_features (number of features for split nodes), and n_estimator (number of trees built). Model evaluation was done using stratified K-fold cross-validation with K = 10. The evaluation metrics used were as follows (Equation 1-4).

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{4}$$

Where A: accuracy, R: recall, P: precision, TP: true positive, TN: true negative, FP: false positive, and FN: false negative.

Results and Discussions

Data Acquisition

There are a total of 113 proteins on the OMIM and Uniprot sites and literature studies search results. For drug-target interaction data, from the 113 proteins, only 36 proteins have interactions with a compound on the DrugBank and SuperTarget sites. The searching process for compound-protein interaction data on the Drugbank site resulted in 786 interactions while the SuperTarget site resulted in 3415 interactions.

Data Preprocess

The data merging process from both sites and removing redundant data process resulted in a total interaction data of 4090 interactions. The feature extraction process on compounds using PubChem fingerprint resulted in 881 attributes. Feature extraction with amino acid composition on protein resulted in 20 attributes. The total attributes in the dataset after the feature extraction process are 904 attributes. The final dataset resulted from data transformation consists of 142488 interactions with 4090 positive interactions and 138398 unknown interactions.

DTI Prediction Model and Evaluation

The feature selection process with XGBoost algorithm reduces the data dimensions from 904 features to 451 features. Random undersampling on a dataset is done by collecting all row data with positive interactions and randomly selects 20450 row data with unknown interactions. The model is then built using a random forest algorithm. Another prediction model for DTI was also built using LGBM, GBM, and XGBoost models to compare the results of random forest performance. All model is built using default parameter and evaluated using stratified K-fold cross-validation with K=10. The comparison of the model results can be seen in Table 2.

Table 2. Data examples

Metrics	Models			
	RF	LGBM	GBM	XGB
Accuracy	0.979±0.002	0.973±0.003	0.942±0.003	0.976±0.003
Recall	0.919±0.010	0.917±0.013	0.825±0.013	0.926±0.009
Precision	0.953±0.008	0.922±0.010	0.828±0.017	0.934±0.012
F-measure	0.936±0.007	0.919±0.008	0.827±0.006	0.930±0.009
AUROC	0.955±0.005	0.950±0.007	0.895±0.005	0.956±0.005

Furthermore, parameter tuning is carried out to try to improve the performance of the RF model using GridSearchCV with K = 5.

The optimal parameter obtained are max_depth = 30, max_features = 32 and n_estimator = 200. Random forest model with optimal parameter produced an average accuracy value of 0.980±0.002, recall value of 0.925±0.008, precision value of 0.954±0.006, f-measure value of 0.939±0.005, and AUROC value of 0.958±0.004. There is an increase in performance compared to the RF model before parameter tuning. Furthermore, this model will be used to predict the dataset before the random undersampling process.

Discussion

There are a total of 113 proteins on the OMIM and Uniprot sites and literature studies search results. 7 proteins are related to COVID-19 in the human body according to OMIM site, 28 COVID-19-related protein according to Uniprot, 19 COVID-19-related protein, and 49 protein that has similar characteristics to ACE2 protein which protein that associated with COVID-19 according to Kuleshov et al. [12]. Kermali et al. also stated that 7 different proteins have associations to COVID-19 [13]. In Zhang and Guo. (2020) and Chen et al., (2020) research, there is 1 protein each associated with COVID-19, whereas in the study of Coleman et al. there is 1 protein associated with the SARS-CoV virus. The total protein data used is 113 proteins [14-16]. Compound-protein interaction network can be seen in Figure 3.

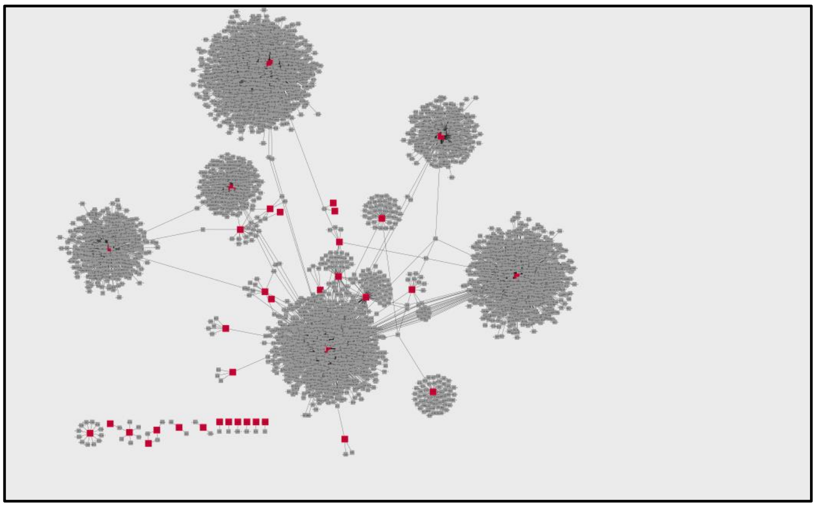

Figure 3. Compound-protein interaction network. Red nodes indicate protein while gray nodes indicate compound

Table 2 shows that the random forest model has better performance on the accuracy, precision, and f-measure metrics compared to other models. Only the XGB model has better performance on precision and AUROC. All models have a standard deviation that tends to be low, which indicates that the performance of all models is quite stable for each fold on the CV.

From these results the random forest model has good performance in predicting DTI (high accuracy), a good prediction of positive class (high recall), has good positive prediction (high precision), has good performance in minority class (f -measure high) and able to distinguish positive and unknown classes well (high AUROC).

Conclusion

In this research, a DTI prediction model using random forest has been implemented. The random forest model built had a good performance in predicting DTI. The random forest produced an average accuracy of 0.979, recall value of 0.919, precision value of 0.953, f-measure value of 0.936, and AUROC value of 0.955. In addition, the random forest also produced an accuracy value of 0.992, recall value of 0.933, precision value of 0.947, f-measure value of 0.94, and AUROC value of 0.992 when used to predict the initial dataset before the undersampling process. For further research, a parameter tuning process with more parameters and a wider interval can be conducted to determine the optimal parameter value. Other scenarios can also be done to solve the problem of class imbalance in the dataset such as SMOTE, oversampling, etc.

Acknowledgement

This research was supported by Ministry of Research, Technology and Higher Education, Indonesia, under Competitive Research Grant from Directorate of Higher Education, Indonesia, 2020 (Contract number: 4020/IT3.L1/PN/2020). The authors also sincerely thank to Department of Computer Science, IPB Universitas for giving support to accomplish the research.

References

1. Malik YS, Shubhankar S, Bhat S, Sharun K, Dhama K, Dadar M, et al. (2020) Emerging novel Coronavirus (2019-nCoV) - Current scenario, evolutionary perspective based on genome analysis and recent developments. Vet. Q. 40(1):68-76. https://doi.org/10.1080/01652176.2020.1727993
2. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, Wang Q, et al. (2020) Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B. 10(5):766-88. https://doi.org/10.1016/j.apsb.2020.02.008
3. Yadav, M, Dhagat, S, Eswari, JS. Emerging strategies on in silico drug development against COVID-19: challenges and opportunities. Eur J Pharm Sci. 2020;155(1):1-15. https://doi.org/10.1016/j.ejps.2020.105522

4. Huang, F, Zhang, C, Liu, Q, Zhao, Y, Zhang, Y, Qin, Y, et al. (2020) Identification of amitriptyline HCl, flavin adenine dinucleotide, azacitidine and calcitriol as repurposing drugs for influenza A H5N1 virus-induced lung injury. PLoS Pathog. 2020;16(3):e1008341. https://doi.org/10.1371/journal.ppat.1008341
5. Sulistiawan, F, Kusuma, WA, Ramadhanti, N, Tedjo, A (2020) Drug-target interaction prediction in coronavirus disease 2019 case using deep semi-supervised learning model. The 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2020); 2020 October 17-18. Indonesia: Universitas Indonesia https://doi.org/10.1109/ICACSIS51025.2020.9263241
6. Mahmud SMH, Chen W, Meng H, Jahan H, Liu Y, Hasan SMM (2019) Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. Anal Biochem 589, 113507.https://doi.org/10.1016/j.ab.2019.113507
7. Spelmen VS, Porkodi R (2018). A Review on Handling Imbalanced Data. Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018: 1-3 March 2018; Coimbatore, India. India: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICCTCT.2018.8551020
8. Chu Y, Shan X, Chen T, Jiang M, Wang Y, Wang Q, Salahub DR, Xiong Y, Wei DQ (2020) DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. Briefings in Bioinformatics 22(3), bbaa205. https://doi.org/10.1093/bib/bbaa205
9. Gortari EF, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): An approach to represent molecular databases. J Cheminform. 2017; 9(9):1-9. doi:10.1186/s13321-017-0195-1.
10. Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 29(2):97–101. https://doi.org/10.1021/ci00062a008
11. Bhasin M, Raghava GPS (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 279(22): 23262- 23266. https://doi.org/10.1074/jbc.M401932200
12. Kuleshov MV, Clarke D, Kropiwnicki E, Jagodnik KM, Bartal A, Evangelista JE, et al. (2020). The COVID-19 Gene and Drug Set Library. Res Sq. 2020;rs.3.rs-28582. https://doi.org/10.21203/rs.3.rs-28582/v1
13. Kermali M, Khalsa RK, Pillai K, Ismail Z, Harky A (2020) The role of biomarkers in diagnosis of COVID-19 - A systematic review. Life Sci 254(117788):1-12. https://doi.org/10.1016/j.lfs.2020.117788
14. Zhang L, Guo H (2020) Biomarkers of COVID-19 and technologies to combat SARS-CoV-2. Adv Biomark Sci Technol 2:1-23. https://doi.org/10.1016/j.abst.2020.08.001
15. Chen X, Yin YH, Zhang MY, Liu JY, Li R, Qu YQ (2020) Investigating the mechanism of ShuFeng JieDu capsule for the treatment of novel Coronavirus pneumonia (COVID-19) based on network pharmacology. Int J Med Sci 17(16):2511-30. https://doi.org/10.7150/ijms.46378.
16. Coleman CM, Sisk JM, Mingo RM, Nelson EA, White JM, Frieman MB (2016) Abelson kinase inhibitors are potent inhibitors of severe acute respiratory syndrome coronavirus and middle east respiratory syndrome coronavirus fusion. J Virol 90(19):8924–33. https://doi.org/10.1128/JVI.01429-16.