Method Article

Progress in Drug Discovery & Biomedical Science

HH PUBLISHER

An automated workflow by using KNIME Analytical Platform: a case study for modelling and predicting HIV-1 protease inhibitors

Ramtin Ranji[1,3], Chanat Thanavanich[2], Sri Devi Sukumaran[1], Sila Kittiwachana[2], Sharifuddin Md. Zain[1], Liew Chee Sun[3], Vannajan Sanghiran Lee[1,*]

[1]Department of Chemistry, Drug Design and Development Research Group (DDDRG), Centre for Theoretical and Computational Physics, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
[2]Department of Chemistry, Faculty of Science, Chiang Mai University, 50200, Chiang Mai, Thailand
[3]Department of Computer System and Technology, University of Malaya, Faculty of Computer Science and Information Technology, 50603, Kuala Lumpur, Malaysia

Abstract: In this study, we have demonstrated an automated workflow by using KNIME Analytical Platform for modelling and predicting potential HIV-1 protease (HIVP) inhibitors. The workflow has been simplified in three easy steps i.e., 1) retrieve the database of inhibitors for the target disease from ChEMBL website and well-known drug from DrugBank database, 2) generate the descriptors and, 3) select the optimal number of features after machine learning models training. Our results have indicated that the random forest with auto prediction validation method is the most reliable with the best R^2 value of 0.9394. Apparently, this workflow can be transformed easily for any other diseases and the quantitative structure-activity relationship (QSAR) model that has been developed can accurately predict *in silico* how chemical modifications might influence biological behaviour. Overall, the automated workflow which has been presented in this study may significantly reduce the time, cost and efforts needed to design or develop potential HIVP inhibitors.

Keywords: Automated workflow; ChEMBL; DrugBank; QSAR; Machine learning, HIV-1 protease inhibitors

Received: 27th August 2019
Accepted: 26th September 2019
Publish Online: 30th September 2019

***Corresponding:** Vannajan Sanghiran Lee, Department of Chemistry, Drug Design and Development Research Group (DDDRG), Centre for Theoretical and Computational Physics, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia; vannajan@um.edu.my

Citation: Ranji R, Thanavanich C, Sukumaran SD, et al. An automated workflow by using KNIME analytical platform: A case study for modelling and predicting HIV-1 protease inhibitors. Prog Drug Discov Biomed Sci 2019; 2(1): a0000035

Introduction

The Human immunodeficiency virus type 1 (HIV-1), has been distinguished as a crucial agent which contributes to the most life-threatening disease, Acquired immunodeficiency syndrome (AIDS)[1]. AIDS is considered as one of the greatest public health and social problems threatening the human race which can be transmitted through various ways i.e., unprotected sexual activities, blood donations without laboratory tests and from mother to children. According to the "The Joint United Nations Programme on HIV/AIDS (UNAIDS)", there were approximately 36.9 million people worldwide living with HIV/AIDS globally in 2017[2]. In 2018, approximately 43% are women and there were about 940,000 deaths from AIDS in 2017[2]. Moreover, the UNAIDS is committed to end the public health threat of the global HIV epidemic by 2030[3]. To achieve this aim, an estimated budget of US$ 26.2 billion will be required for the HIV response in 2020, which may gradually reduce to $22.3 billion by 2030[4].

Studies have shown that the life cycle of the HIV-1 is truly depends on the key enzyme, HIV-1 protease (HIVP)[5, 6]. HIVP cleaves Gag and Gag-Pol polyprotein precursor encoded by the HIV-1 virus genome at nine processing sites to produce mature active proteins[6]. However, the HIVP enzyme activity can be inhibited by blocking the active site of the enzyme by protease inhibitors[6], which results in the release of structurally disorganized and non-infectious viral particles[7]. Therefore, inhibition of HIVP is considered among the most important approaches for the therapeutic intervention in HIV infection[8].

Although there is currently exists no treatment to eradicate the virus from an infected patient, a range of medications can control the condition. To date, 26 anti-HIV compounds have been approved by the Food and Drug Administration (FDA)[9], in which 10 are HIV protease inhibitors[10]. Nevertheless, the treatment of HIV with the FDA-approved drugs is only effective in reducing viral replication, and HIV rapidly gains resistance to all known agents[11]. Hence, there is an urge to look for new drugs to treat the current disease.

Recently, huge amount of experimental data which aids

Copyright © 2019 by Ranji R *et al.* and HH Publisher. This work under licensed under the Creative Commons Attribution-NonCommercial 4.0 International Lisence (CC-BY-NC 4.0)

researchers throughout the drug discovery process become available through open-source websites such as ChEMBL (https://www.ebi.ac.uk/chembl/)[12]. However, flexible analysis of data and the development of scripts which are specific to each project require high skills in software engineering. To overcome this issue, workflow environments[13] such as Pipeline Pilot[14], Taverna[15], Kepler[16], Galaxy[17], Loni Pipeline[18] and KNIME (Konstanz Information Miner)[14, 19] have been emerged for flexible and easy data analysis. In this study, KNIME Analytical Platform will be explored to develop an automated workflow for HIVP inhibitors database and clustering analysis. The quantitative structure-activity relationship (QSAR) model from this workflow can be used to automatically predict the compound activities. Hence, the process of drug discovery against HIVP target is expected to be simplified by the help of automated compound activity prediction.

Method details

Materials

Softwares & databases (ChEMBL, DrugBank, KNIME Analytics Platform and RDKit)

Procedure

Data processing for this study was conducted using the KNIME Analytics Platform (https://www.knime.com/), an open-source software which is used widely in data science to automate the data science process. KNIME is capable of performing all steps required for data analysis in a user-friendly environment for non-experts in software engineering. Each workflow in KNIME consists of many nodes in which each of them performs a certain job. These nodes are built-in or developed by the community. Additionally, the KNIME contains a wide range of community nodes for the analysis of chemical structures such as chemical similarity check.

The workflow of current study consists of four stages i.e., 1) data collection, 2) data clustering, 3) model development and training, and 4) bioactivity prediction and score reviewing. **Figure 1** shows the workflow for data collection. The required bioactivity data targeting HIVP has been downloaded from ChEMBL database (ChEMBL ID: ChEMBL243[20]) and the data were pre-processed. Then, descriptors for each molecule entry in the table were calculated by the help of RDKit nodes (http://www.rdkit.org/), an open source toolkit. Generally, RDkit adds some cheminformatics functionalities such as molecule I/O, substructure searching, and chemical reactions into KNIME.

Next, the data clustering was performed by using the K-means clustering method[21] as shown in **Figure 2**. Normally, K-means clusters data into k number of clusters in which each observation belongs to the cluster that has the nearest mean. In the present study, the data clustering was conducted based on the structure similarity with the Darunavir (DVR), an oral nonpeptidic HIVP inhibitor[22]. Previous study has reported that DVR has a high genetic barrier to resistance and active against multidrug-resistant HIV isolates[22]. Therefore, this drug has been incorporated into data clustering procedure.

Following data clustering, three machine learning models were developed and trained i.e., linear, polynomial and random forest regression with 67% of available data. **Figure 3** shows the random forest learner as a sample of developed models. After exhaustive model training, the optimal number of features should be selected in 'Feature Selection Filter" node. Finally, the trained models were tested with the data that have not been used in the training to validate the trained model. The reusable workflow has been tested for KNIME 3.6 and can be requested from Computational Chemistry Research Group Services, (https://compchem.dicc.um.edu.my/services), Email:compchem@um.edu.my.

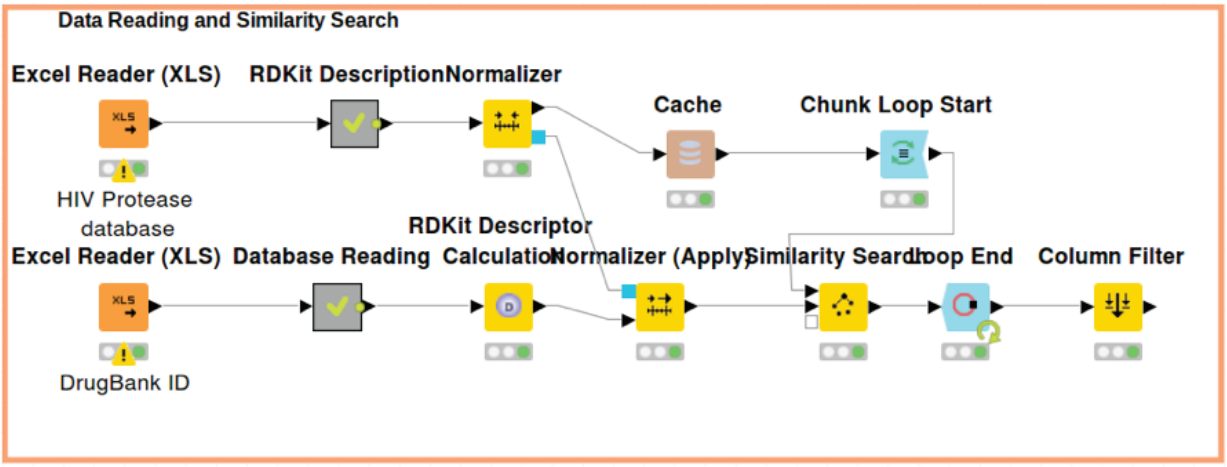


Figure 1. Workflow of reading data from ChEMBL and Drugbank.

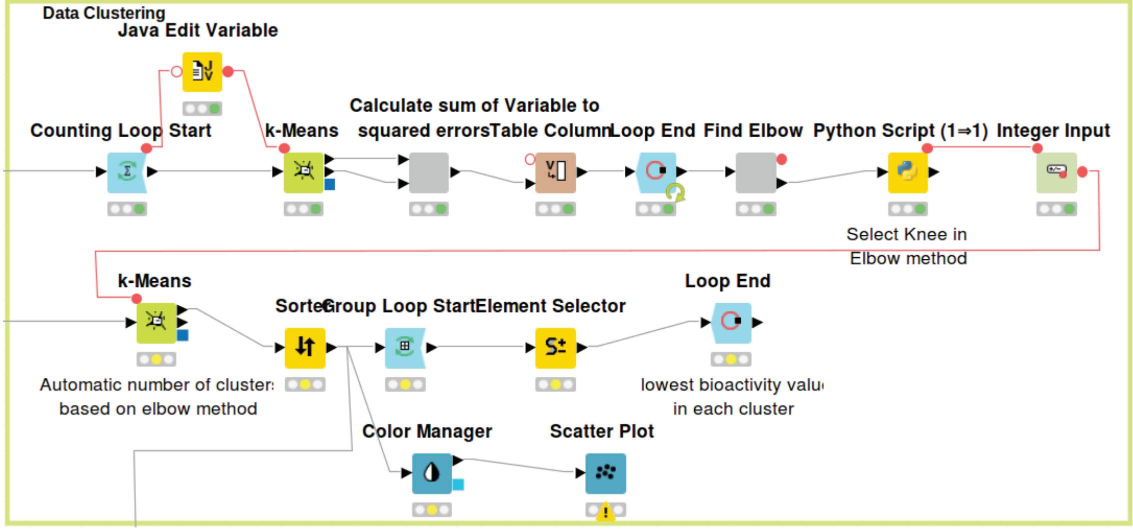


Figure 2. Data clustering workflow based on the similarity of structures and drug.

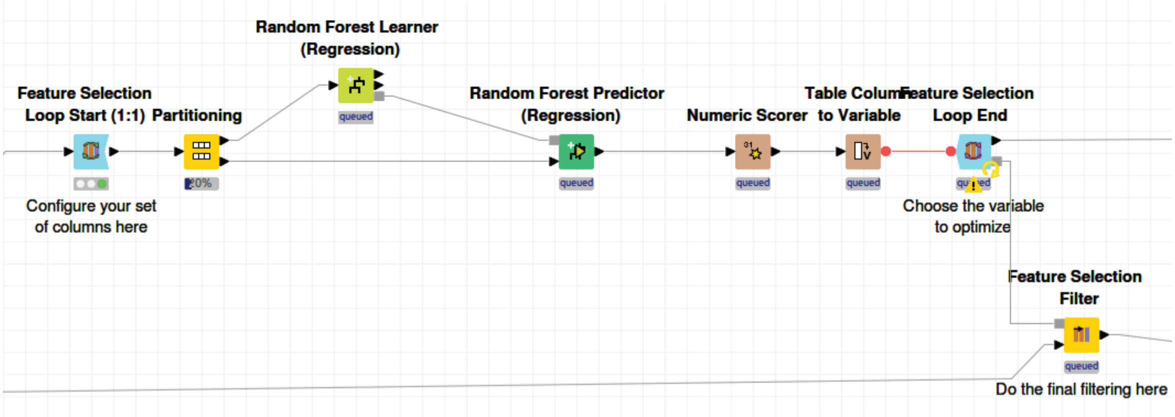
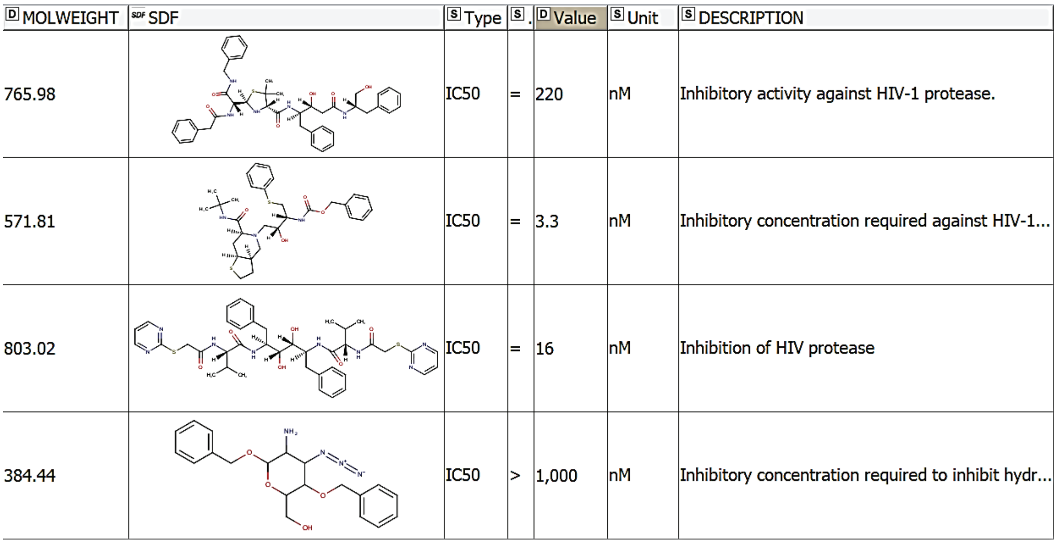


Figure 3. Random forest machine learning model training to find the optimal features.

Results and discussion

During data reading process, a total of 8241 data were extracted from ChEMBL database[20] in which 2069 were selected based on the inclusion of IC_{50} value. The DVR was retrieved from the DrugBank database (https://www.drugbank.ca/) (DrugBank ID: DB01264). The examples of data which have been downloaded and calculated the descriptors by RDKit are shown in **Figure 4**. For data clustering, the sum of squared errors were calculated with K-means clustering followed by determination of the best number of clusters using elbow method. The results have indicated that the data must be clustered into three groups (**Figure 5**). Subsequently, three clusters were identified and the highest bioactive structure has been shown in each cluster (**Figure 6**).



D MOLWEIGHT	SDF SDF	S Type	S .	D Value	S Unit	S DESCRIPTION
765.98		IC50	=	220	nM	Inhibitory activity against HIV-1 protease.
571.81		IC50	=	3.3	nM	Inhibitory concentration required against HIV-1...
803.02		IC50	=	16	nM	Inhibition of HIV protease
384.44		IC50	>	1,000	nM	Inhibitory concentration required to inhibit hydr...

Figure 4. Samples of fetched data.

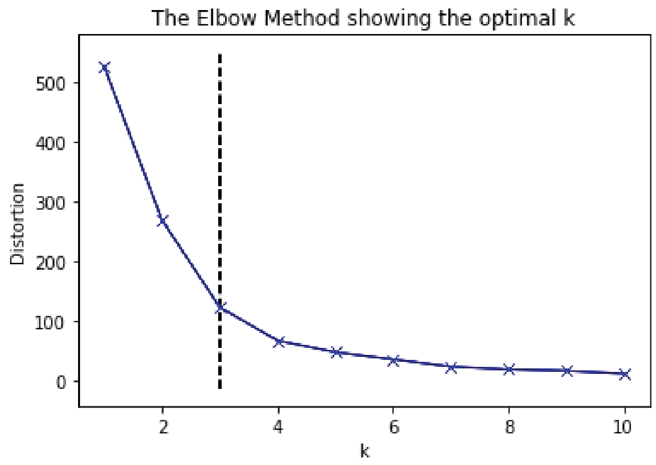


Figure 5. Optimal number of clusters based on the elbow method in K-means structural clustering.

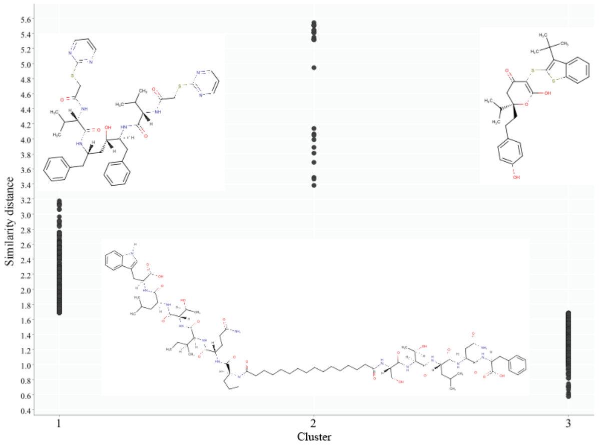


Figure 6. Clusters based on the similarity distance from Darunavir with the structural diversity. The best activity models in each cluster are shown.

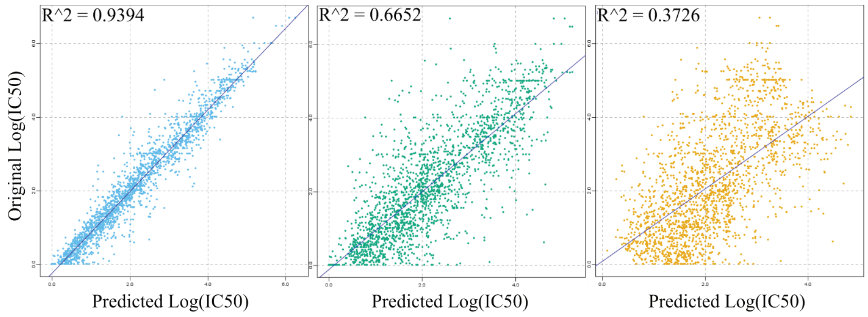
In training and validation steps, the automated activity prediction has been performed using three different learning methods i.e., random forest, linear and polynomial regression. As shown in **Table 1**, each method has achieved different performance which represents by different value of coefficient of determination, R^2. Moreover, the trained models were evaluated by 10-fold and leave-one-out cross validation (CV) methods.

V-fold cross-validation (VFCV) is a very popular method due to its low computational cost. In this method, the size of the training set is $n_t = n(V-1)/V$ and n/V of data are used for validation purpose. Furthermore, the reliability of a model is largely depends on the value of V. Many studies have suggested that if the value of V=10, the model is more accurate[23]. In contrast, Leave-one-out cross validation (LOOCV) technique is computationally intensive[24]. In this CV method, $n_t = n-1$ which means that in each iteration only one of the data points left out for validation.

Our results have indicated that the random forest with auto prediction validation method is the most reliable with the best R^2 value of 0.9394. The bioactivity prediction with random forest but different validation methods is depicted in **Figure 7**.

Table 1. Results of $LogIC_{50}$ value prediction by different methods.

Learning Method	Validation Method	Structures	Coefficient of determination (R^2)
Linear regression	Auto Prediction	All	0.474
Linear regression	Leave one out	All	-2.368
Linear regression	10-fold	All	-1.122
Polynomial regression	Auto Prediction	All	0.167
Polynomial regression	Leave one out	All	0.158
Polynomial regression	10-fold	All	-0.045
Random forest	**Auto Prediction**	**All**	**0.939**
Random forest	Leave one out	All	0.663
Random forest	10-fold	All	0.371



A) Auto Prediction B) Leave-1-out cross validation C) 10-Fold cross validation

Figure 7. Prediction of $logIC_{50}$ values with random forest regression.

Conclusion

In this paper, we have presented an automated workflow by using KNIME Analytical Platform for HIVP inhibitors. A QSAR model has been developed to predict the bioactivity features of the compounds. Furthermore, this workflow can be adapted for other diseases in three easy steps i.e., 1) automatically download the database of inhibitors for the target disease from ChEBML website and the well-known drug from DrugBank database, 2) generate the descriptors 3) select the optimal number of features for machine learning models training and prediction. Therefore, this study paved a way to simpler process of drug discovery.

Conflict of Interest

The authors declare that there is no conflict of interest.

Acknowledgements

The authors thank the University of Malaya for providing grants (RP037A-17AFR) to conduct this study.

Reference

1. Buonaguro L, Tornesello ML, and Buonaguro FM, Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: Pathogenetic and therapeutic implications. *Journal of Virology*, 2007. 81(19): 10209-10219.
2. “Fact Sheet” (PDF). UNAIDS.org. 2018.
3. “UNAIDS Strategy,” http://www.unaids.org/en/goals/unaidsstrategy.
4. “Homepage,” http://www.unaids.org/en/Homepage.
5. Zhang S, Kaplan AH, and Tropsha A, HIV-1 protease function and structure studies with the simplicial neighborhood analysis of protein packing method. *Proteins*, 2008. 73(3): 742-753.
6. Lv Z, Chu Y,Wang Y. HIV protease inhibitors: a review of molecular selectivity and toxicity. HIV/AIDS. Auckland, N.Z; 2015. p. 95–104.
7. Temesgen Z, and Wright AJ, Recent advances in the management of human immunodeficiency virus infection. *Mayo Clinic Proceedings*, 1997. 72(9): 854-858.
8. Mudgal M, Birudukota N, and Doke M, Applications of Click Chemistry in the Development of HIV Protease Inhibitors. *International Journal of Medicinal Chemistry*, 2018. 2018: 9 pages.
9. Win NN, Ngwe H, Abe I, *et al*., Naturally occurring Vpr inhibitors from medicinal plants of myanmar. *Journal of Natural Medicines*, 2017. 71(4): 579-589.
10. Humpolíčková J, Weber J, Starková J, *et al*., Inhibition of the precursor and mature forms of HIV-1 protease as a tool for drug evaluation. *Scientific Reports*, 2018. 8(1): 10438.
11. Richter SN, Frasson I, and Palù G, Strategies for inhibiting function of HIV-1 accessory proteins: a necessary route to AIDS therapy?. *Current Medicinal Chemistry*, 2009. 16: 267-286.
12. Gaulton A, Bellis LJ, Bento AP, *et al*., ChEMBl: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 2012. 40(Database issue): D1100-D1107.
13. Nicola G, Berthold M, Hedrick M, *et al*., Connecting proteins with drug-like compounds: Open source drug discovery workflows with BindingDB and KNIME. *Database*, 2015. 2015: bav087.
14. Warr WA, Scientific workflow systems: Pipeline pilot and KNIME. *Journal of Computer-Aided Molecular Design*, 2012. 26(7): 801-804.
15. Wolstencroft K, Haines R, Fellows D, *et al*., The taverna workflow suite: Designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Research*, 2013. 41(W1): W557-W561.
16. Ludäscher B, Altintas I, Berkley C, *et al*., Scientific workflow management and the kepler system: Research articles. *Concurrency and Computation*: *Practice* & *Experience*, 2006. 18(10): 1039-1065.
17. Afgan E, Baker D, Batut B, *et al*., The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 2018. 46(W1): W537-W544.
18. Rex DE, Ma JQ, and Toga AW, The loni pipeline processing environment. *Neuroimage*, 2003. 19(3): 1033-1048.
19. Fillbrunn A, Dietz C, Pfeuffer J, *et al*., Knime for reproducible cross-domain analysis of life science data. *Journal of Biotechnology*, 2017. 61: 149-156.
20. ChEMBL Database. (2019). Retrieved 10 May 2019, from https://www.ebi.ac.uk/chembl/g/#browse/activities/filter/target_chembl_id%3ACHEMBL243
21. Jin X, Han J. K-means clustering. In C. Sammut & G. I. Webb (Eds.), Encyclopedia of machine learning. Boston, MA: Springer US; 2010. p. 563-564.
22. McKeage K, Perry C, and Keam S, Darunavir. *Drugs*, 2009. *69*(4): 477-503.
23. Arlot S, and Celisse A, A survey of cross-validation procedures for model selection. *Statistics Surveys*, 2010. 4(0): 40-79.
24. Cheng H, Garrick DJ, and Fernando RL, Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of Animal Science and Biotechnology*, 2017. 8: 38.