MDPI sciforum

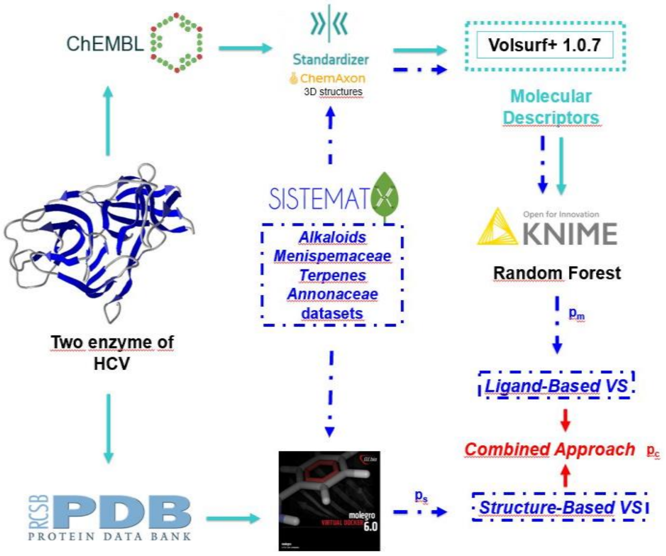
MOL2NET, International Conference Series on Multidisciplinary Sciences

Ligand-Based and Structure-Based virtual screening for the discovery of natural inhibitor the Hepatitis C Virus

<Renata Priscila Costa Barros> (renatabarros@ltf.ufpb.br)[a], <Marcus Tullius Scotti> (mtscotti@gmail.com)[a]

[a] *< Post-Graduate Program in Natural Synthetic Bioactive Products, Federal university of Paraiba >*

Graphical Abstract



Abstract. Hepatitis C is a disease that constitutes a serious global health problem, is often asymptomatic and difficult to diagnose, about 60-80% of infected patients develop chronic diseases over time. As there is no vaccine against hepatitis C virus (HCV), developing new cheap treatments is a big challenge. The search for new drugs from natural products has been outstanding in recent years. The aim of this study was combining structure-based and ligand-based virtual screening (VS) techniques to select potentially active molecules against two HCV target proteins from in-house secondary metabolite dataset (SistematX). From the ChEMBL database, we selected two sets of 1199 and 237 chemical structures with inhibitory activity against different targets of HCV to create random forest models with an accuracy value higher than 72% for cross-validation and test sets. Afterward, a ligand-based virtual screen of the entire 1378 secondary metabolites database stored in SistematX was performed. In addition, a structure-based virtual screening was also performed for the same set of secondary metabolites using molecular docking. Finally, using consensus analyzes approach combining ligand-based and structure-based VS, two alkaloids and one triterpene were selected as potential anti-HCV compounds.

Introduction

Hepatitis C is a disease that constituted a serious public health problem worldwide, is a liver disease and makes patients affected by this disease vulnerable to cirrhosis, hepatic insufficiency, hepatocellular carcinoma, etc. [1]. Hepatitis C has two stages, the acute phase that is difficult to diagnose at the onset of infection, and the chronic phase [2].

A new era in the treatment of hepatitis C was started in 2011 from the approval of first generation direct acting antiviral agents (AADs) [3]. In 2014, more potent and better new AADs became available by achieving a sustained immune response (SVR). However, several studies have shown that patients with hepatic cirrhosis tend to have early HCV RNA recurrence in weeks and recur after cessation of treatment with AADs. A major problem in this new era of HCV treatment is the extremely high cost of medications [3, 4].

Due to the high efficiency, ie low cost of research, decreased exploratory testing and better understanding and interpretation of biochemical systems and therapeutic targets, computational techniques have been widely used [5].

In this perspective, a combination of ligand-based and virtual structure-based screening techniques was performed on secondary metabolite banks (Menispermaceae alkaloids and Annonaceae terpenes) to select the best active molecules against two HCV target proteins.

Materials and Methods

1. Dataset

From the ChEMBL database, we selected two sets of chemical structures with inhibitory activity against different targets of hepatitis C virus for construction of predictive models. The compounds were classified using values of –logIC50 (mol/L) = pIC50. In this case, IC50 represented the concentration required for 50% inhibition of enzymes of HCV.

Two datasets of secondary metabolites composed by 809 Menispermaceae alkaloids and 569 Annonaceae terpenes were extracted from our in-house databank SistematX available at http://sistematx.ufpb.br [6]. These databases were used for virtual screening to select the molecules with the highest values of probability to inhibit the HCV targets. For all structures, SMILES codes were used as input data in Marvin 18.10.0, 2018, ChemAxon (http://www.chemaxon.com). We used Standardizer software JChem17.29.0, 2017; ChemAxon (http://www.chemaxon.com)] to canonize structures, add hydrogens, perform aromatic form conversions, clean the molecular graph in three dimensions and save compounds in sdf format.

Three-dimensional (3-D) structures were used as input data in the Volsurf+ program v. 1.0.7 [7] and were subjected to molecular interaction fields (MIFs) to generate descriptors a total of 128 descriptors.

1.2 Prediction Model

The Knime 3.4.0 software (Knime 3.4.0 the Konstanz Information Miner Copyright, 2003–2014, www.knime.org) was used to perform all of the following analyses. The descriptors and class variables were imported from the software Volsurf+ program v. 1.0.7, and for each one the data were divided using the “partitioning” node with the “stratified sample” option to create a training set and a test set, encompassing 80% and 20% of the compounds, respectively. Although the compounds were selected randomly, the same proportion of active and inactive samples was maintained in both sets.

For internal validation, we employed cross-validation using 10 randomly selected, stratified groups, and the distributions according to activity class variables were found to be maintained in all validation groups and in the training set. Descriptors were selected, and a model was generated using the training set and the Random Forest algorithm (RF), using the WEKA nodes [8, 9].

The internal and external performances of the selected models were analyzed for sensitivity (true positive rate, i.e., active rate), specificity (true negative rate, i.e., inactive rate) and accuracy (overall predictability). In addition, the sensitivity and specificity of the Receiver Operating Characteristic (ROC) curve were found to describe true performance with more clarity than accuracy. Using Knime nodes the most important descriptors in the generation of each prediction model were evaluated.

The domain of applicability (APD) was used to analyze the compounds of the test sets evaluating whether or not their predictions were reliable. The APD is based on Euclidean distances and similarity measures between the descriptors of the training set are used to define the applicability domain, so if a test set compound has distances and similarity beyond this limit, its prediction is not reliable [10].

1.3 Molecular Docking

The structures of two HCV proteins, the NS3 protesse type 1 (ID 3EYD) [11] and Polymerase NS5B (ID 3H5S) [12] in a complex with the respective inhibitors were downloaded from the Protein Data Bank (PDB), each protein referring to an HCV model. All water molecules were deleted from the enzyme structure, and the enzyme and compound structures were prepared using the same default parameter settings in the same software package (score function: MolDock Score; ligand evaluation: internal ES, internal H-bond, sp2–sp2 torsions, all checked; number of runs: 10 runs; algorithm: MolDock SE; maximum interactions: 1500; max. population size: 50; max. steps: 300; neighbor distance factor: 1.00; max. number of conformations returned: 5). The docking procedure was performed using a GRID with a radius of 15 Å and a resolution of 0.30 Å to cover the ligand-binding site in the structures of the two enzymes.

Results and Discussion

The models had the positive interest rates were lower than 28%, with access rates higher than 72%. Two parameters were used to evaluate the quality of these binary models: The Receiver Operating Characteristic (ROC) curve and Matthews correlation coefficient (MCC) [13,14]. In both models, the area under the curve was greater than 81% for the cross-validation sets, and greater than 85% for the test sets, revealing that the models are capable of performing a good classification and prediction rate. Figure 1 shows the ROC curves of the test and cross-validation for all models.

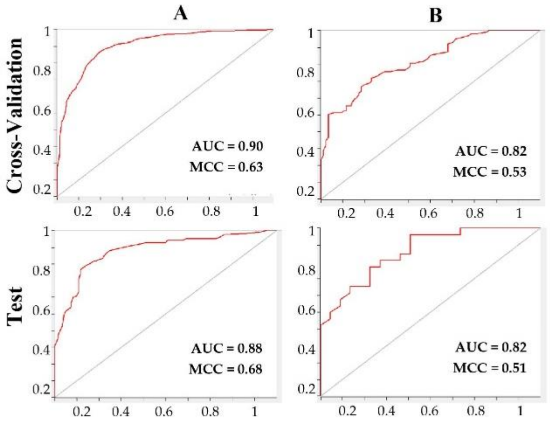


Figure 1: ROC plot, sensitivity versus 1-specificity, generated for the selected RF model for cross-validation and test sets: A) NS5B RNA dependent RNA polymerase, B) Polyprotein - genotype 1a. AUC = value of the area under the curve; MCC = Matthews Correlation Coefficient.

MCC values for training, cross-validation and test sets in RF models were obtained. For both models, a perfect correlation (MCC = 1) was observed for the training set, with the NS5B RNA dependent RNA polymerase model presenting the highest values 0.78 (cross-validation) and 0.79 (test). All MCC values are shown in Figure 1.

The molecular docking, ligand-based virtual screening, was first validated by redocking of the original ligand for each of the two HCV proteins used in this study. The MolDock scores are listed in Table 1 along with their respective RMSD values and the energies from the PDB.

Table 1: The docking energy (kJ/mol) of the ligand PDB for each of the two HCV enzymes. Ligand energy of the MolDock score and the RMSD values obtained from the redocking procedure.

Proteins	Energy PDB (KJ/mol)	Energy Moldock (KJ/mol)	Redocking RMSD
3EYD	-141.164	-143.059	0.99
3H5S	-159.014	-171.225	0.21

Therefore, using the same parameters for all proteins, a virtual screening for all molecules of reliable secondary metabolites of each prediction model was realized . Based on the binding energy values, all tested molecules were ranked using the following probability calculation (Equation 1):

$$\text{ps} = \frac{E_{TM}}{E_M}, IF\ E_{TM} < E_L \qquad (1)$$

where *p*s = structure-based probability, E_{TM} = docking energy of molecule test and TM ranges from 1 to 1848 (secondary metabolites dataset); Emin = the average value of the energies of the molecules of the dataset; Eligand = the ligand energy from protein crystallography.

The secondary metabolites were classified as active if the structure-based probability values are greater than or equal to 0.5. The numbers of molecules with probability values greater than 0.5 and binding energy values less than the ligand were 3EYD (140) and 3H5S (558).

An approach combining structure-based and ligand-based virtual screening was realized to verify potentially active molecules as well as their possible mechanism of action, showing potential multitarget molecules. This approach also seeks to minimize the probability of selecting false positive molecules because it considers the scores of both virtual tracking techniques and correlates them with the true negative rate [15,16]. The calculation is done with the following equation:

$$P_c = \frac{p_s+(1+TN)\ x\ p}{2+TN} \qquad (2)$$

where P_c is combined probability, p_s is the structure-based probability, TN is the true negative rate and p is the ligand-based probability. In this equation, the ligand-based score is conditioned to a decrease in the false positive rate with the increment of TN. Thus, the probability of selecting inactive molecules as active molecules is minimized. Table 2 summarizes the results for the best-ranked molecules obtained using the combined approach.

Table 2: Summary of the best-ranked structures obtained using an approach combining ligand-based and structure-based virtual screening; p = active probability value in ligand-based VS; p_s = active probability value in structure-based VS. P_c = combined probability value.

Protein	Molecule	*p*	p_s	p_c
3EYD	Secocepharanthine	0.59	0.86	0.69
	Dihydroseccocepharanthine	0.48	0.81	0.60
	Cycloart-24-en-3-b-21-b-diyl diacetate, 5-a: 21(R)-23(R)-epoxy:	0.51	0.76	0.60
3H5S	Secocepharanthine	0.62	0.91	0.72
	Dihydroseccocepharanthine	0.82	0.88	0.84
	Cycloart-24-en-3-b-21-b-diyl diacetate, 5-a: 21(R)-23(R)-epoxy:	0.66	0.88	0.73

Conclusions

In this study, we selected three secondary metabolites as potential multitargets anti-HCV through rapid approaches using ligand-based and structure-based VS of 1378 alkaloids and terpenes of three botanical families, Meninspermaceae and Annonaceae, obtained from an in-house database. The compounds selected have structural similarities with other secondary metabolites related in the literature as antiviral compounds. The selected structures are a start point to further studies in order to develop new anti-HCV compounds based on natural products.

References

1. Ismail, N. S. M., Elzahabi, H. S. A., Sabry, P., Baselious, F. N., AbdelMalaK, A. S., Hanna, F. A study of the allosteric inhibition of HCV RNA-dependent RNA polymerase and implementing virtual screening for the selection of promising dual-site inhibitors with low resistance potential. *J Recept Sig Transd*, **2016**, *Volume* 37, p. 341-354, http://dx.doi.org/10.1080/10799893.2016.1248293.

2. Ganesan, A.; Barakat, K. Applications of computer-aided approaches in the development of hepatitis C antiviral agents. *Expert Opin Drug Discov*., **2017**, *Volume* 12, p. 407-425, http://dx.doi.org/10.1080/17460441.2017.1291628.
3. Cheung, M. C., Walker, A. J., Hudson, B. E., Verma, S., Mc Lauchlan, J., Mutimer, D. J., Brown, A., Gelson, W. H. T., MacDonald, D. C., Agarwal, K., Foster, G. R., Irving, W. L., HCV Research UK. Outcomes after successful direct-acting antiviral therapy for patients with chronic hepatitis C and decompensated cirrhosis. J Hepatol, 2016, Volume 65, p. 741–747, http://dx.doi.org/ 10.1016/j.jhep.2016.06.019
4. Foster, G. R., Irving, W. L., Cheung, M. C., Walker, A. J., Hudson, B. E., Verma S., McLauchlan, J., Mutimer, D. J., Brown, A., Gelson, W. T., MacDonald, D. C., Agarwal, K., HCV Research UK. Impact of direct acting antiviral therapy in patients with chronic hepatitis C and decompensated cirrhosis. J Hepatol, 2016, Volume 64, p. 1224–1231, http://dx.doi.org/10.1016/j.jhep.2016.01.029.
5. Ponder, E. L., Freundlich, J. S., Sarker, M., Ekins, s. Computational models for neglected diseases: gaps and opportunities. *Pharm Res*, **2014**, *Volume* 31, n. 2, p. 271-277, https://doi.org/10.1007/s11095-013-1170-9.
6. Scotti, M. T., Herrera-Hacevedo, C., Oliveira, T. B., Costa, R. P. O., Santos, S. Y. K. O., Rodrigues, R. P., Scotti, L., Da-Costa, F. B. SistematX, an Online Web-Based Cheminformatics Tool for Data Management of Secondary Metabolites. *Molecules*, **2018**, *Volume* 23, p. 103-113, http://dx.doi.org/10.3390/molecules23010103
7. Cruciani, G., Crivori, P., Carrupt, P.-A., Testa, B. Predicting Blood−Brain Barrier Permeation from Three-Dimensional Molecular Structure. *J. Mol. Struct*., 2000, *Volume* 503, p. 17-30, http://dx.doi.org/10.1021/jm990968+
8. Steven L. Book Review: C4.5: Programs for Machine Learning. *Morgan Kaufmann*, **1993**, *Volume* 16, p. 235-240, https://doi.org/10.1023/A:1022645310020.
9. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. Witten, I.H. The WEKA data mining software: an update. *SIGKDD Exploration*, **2009**, *Volume* 11, p. 10 -18.
10. Scotti, M. T., Scotti, L., Ishiki, H. M., Peron, L. M., Rezende, L., Amaral, A. T. Variable selection approaches to generate QSAR models for a set of antichagasic semicarbazones and analogues. *Chemom. Intell. Lab. Syst.* **2016**, *Volume* 154, p. 137-149.
11. Venkatraman, S., Wu, W., Prongay, A., Girijavallabhan, V., George Njoroge, F. Potent inhibitors of HCV-NS3 protease derived from boronic acids. *Bioorg Med Chem Lett*, **2009**, *Volume* 19, p. 180-183, http://dx.doi.org/10.1016/j.bmcl.2008.10.124.
12. Vicente, J., Hendricks, R. T., Smith, D. H., Fell, J. B., Fischer, J., Spencer, S. R., Stengel, P. J., Mohr, P., Robinson, J. E., Blake, J. F., Hilgenkamp, R. K., Yee, C., Adjabeng, G., Elworthy, T. R., Li, J., Wang, B., Bamberg, J. T., Harris, S. F., Wong, A., Leveque, V. J., Najera, I., Le Pogam, S., Rajyaguru, S., Ao-leong, G., Alexandrova, L., Larrabee, S., Barndl, m., Briggs, A., Sukhtankar, S., Farrell, R. Non-nucleoside inhibitors of HCV polymerase NS5B. Part 4: Structure-based design, synthesis, and biological evaluation of benzo[d]isothiazole-1,1-dioxides. *Bioorg Med Chem Lett*, **2009**, *Volume* 19, p. 5652-5656, http://dx.doi.org/10.1016/j.bmcl.2009.08.022.
13. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta Prot Struct*, **1975**, *Volume* 405, p. 442-451, http://dx.doi.org/10.1016/0005-2795,(75)90109-9.
14. Silva, F. C. (Federal University of Pernambuco, Recife, Pernambuco, Brazil); Análise ROC, 2006.
15. Lorenzo, V., Lúcio, A. S., Tavares, J. F., Filho, J. M., Lima, T. K., Rocha, J. D., Scotti, M. T. Structure- and Ligand-Based Approaches to Evaluate Aporphynic Alkaloids from Annonaceae as Multi-Target Agent Against Leishmania donovani. *Curr Pharm Des*., **2016**, *Volume* 22, p. 5196-5203, http://dx.doi.org/10.2174/1381612822666160513144853.
16. Acevedo, C. H., Scotti, L.; Scotti, M. T. In Silico Studies Designed to Select Sesquiterpene Lactones with Potential Antichagasic Activity from an In-House Asteraceae Database. ***Chemmedchem***, **2018**, *Volume* 13, p.634-645, http://dx.doi.org/10.1002/cmdc.201700743.