Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 107 (2019) 127–130

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

AntiVPP 1.0: A portable tool for prediction of antiviral peptides

Jorge Félix Beltrán Lissabet, Lisandra Herrera Belén, Jorge G. Farias*

Universidad de La Frontera, Department of Chemical Engineering, Faculty of Engineering and Science, Temuco, Chile

ARTICLE INFO

Keywords:
Peptide
Prediction
Antiviral
Machine learning
Python
Software

ABSTRACT

Viruses are worldwide pathogens with a high impact on the human population. Despite the constant efforts to fight viral infections, there is a need to discover and design new drug candidates. Antiviral peptides are molecules with confirmed activity and constitute excellent alternatives for the treatment of viral infections. In the present study, we developed AntiVPP 1.0, an accurate bioinformatic tool that uses the Random Forest algorithm for antiviral peptide predictions. The model of AntiVPP 1.0 for antiviral peptide predictions uses several features of 1088 peptides for training and validation. During the validation of the model we achieved the TPR = 0.87, SPC = 0.97, ACC = 0.93 and MCC = 0.87 performance measures, which were indicative of a robust model. AntiVPP 1.0 is a fast, accurate and intuitive software focused on the assessment of antiviral peptides candidates. AntiVPP 1.0 is available at https://github.com/bio-coding/AntiVPP.

1. Introduction

Viruses are very old and ubiquitous pathogens, which cause high rates of infection and mortality in the human population [1]. The success of viruses during evolution has been possible due to three general attributes: genetic variation, the variety of forms for their transmission and the efficient way to replicate within their host cells in order to remain in them [2,3]. Due to these attributes, the control of viral diseases throughout history has not been an easy task [4]. In spite of the existence of antiviral drugs, it is necessary to explore novel antiviral compounds in order to control emerging viral pathogens [4,5].

In recent decades, peptides have become increasingly important in the design and delivery of drugs. Research in this regard is focused on the development and refinement of techniques to design and identify synthetic and natural peptides as drug candidates [1,6]. Antiviral peptides (AVPs) are known to fight against various types of viruses and can come from synthetic combinatorial libraries or segments of natural proteins [5,6]. There are different scenarios in which the AVPs have shown activity, e.g. Enfuvirtide (also known as T20), the first peptide inhibitor approved by the FDA against the HIV-1 [7]. Antiviral activity has also been reported for viruses, e.g. Rabies [8], HCV [9], influenza A virus H1N1, H3N2, H5N1, H7N7, H7N9, SARS-CoV and MERS-CoV [10], among others.

Nowadays, there are different databases that contain collections of AVPs, among them: AVPpred [11], APD3 [12], CAMPR3 [13] and HIPdb [14], which constitutes excellent opportunities for the development of computational tools focused on the prediction of these molecules. However, unlike the development of bioinformatics tools in the field of antimicrobial peptides predictions (bacteria, fungi, animal cells) [15], the development of *in silico* tools for the prediction of AVPs is an area that has remained scarcely explored [11]. Currently, there are only three methods for predicting AVPs. The first one is the AVPpred server, which uses a vector support machine (SVM) for its predictions [11]. The second method is based on Random Forest (RF) algorithm and the resulting model of this work showed a better performance in the prediction of AVPs than AVPpred [16]. However, this model has not software to carry out prediction tasks by researchers who are not related to the field of machine learning. The third method, AVP-IC50Pred, was developed by Quresshi and coworkers. AVP-IC50Pred is a regression-based algorithm which uses experimentally proven datasets by employing multiple machine learning algorithms [17]. In this work, we have developed a friendly and portable software based on the RF algorithm for the prediction of AVPs with excellent performance measurements.

2. Materials and methods

2.1. Datasets

To carry out this study, the data set reported by Thakur et al., was selected [11]. For training of the model, the data set T544p + 544n* was used (a total of 1088 peptides). 544p corresponds to a collection of 544 antiviral peptides with experimentally validated activity, while the 544n* are 544 non-experimental negative peptides, which has been

* Corresponding author. Universidad de La Frontera, Ave. Francisco Salazar, 01145, Temuco, Chile.
E-mail address: jorge.farias@ufrontera.cl (J.G. Farias).

https://doi.org/10.1016/j.compbiomed.2019.02.011
Received 12 January 2019; Received in revised form 15 February 2019; Accepted 16 February 2019

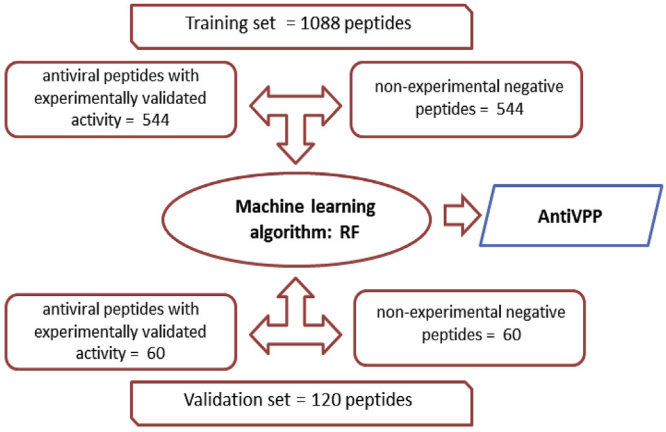

Fig. 1. Architecture of the training and validation model based on the dataset reported by Thakur and coworkers [11].

used in the development of prediction models of antiviral peptides [11,16]. For validation of the model, the independent data set V60p + 60n* was selected, composed of 60 peptides with experimentally validated activity (V60p) and 60 negative non-experimental peptides (60n*) (a total of 120 peptides). The building of the training and validation of the model is shown in Fig. 1.

2.2. Peptide features

For this study, the following features: net charge [18], number of hydrogen bond donors [19], molecular weight [20] and hydropathy index [21], were evaluated. Also, the composition of charged (DEKHR), aliphatic (ILV), aromatic (FHWY), polar (DERKQN), neutral (AGHPSTY), hydrophobic (CVLIMFW), positively charged (HKR), negatively charged (DE), tiny (ACDGST), small (EHILKMNPQV) and large (FRWY) residues as well as the relative frequency of all 20 natural amino acids, were assessed. All features were computed by using the Python 3.6 programming language (available at https://www.python.org/).

2.2.1. Relative frequency (Rfre) of all 20 natural amino acids

$$\text{Rfre}[a.\ a] = Xi/N$$

where Rfre [a.a] is the relative frequency of a natural amino acid of type *i*. N is the total number of natural amino acids in the peptide (peptide length).

2.2.2. Residues composition of peptides (PEP [comp])

Ex: PEP[positively charged] = Rfre[H] + Rfre[K] + Rfre[R]

where PEP [comp] is the sum of all Rfre [a.a] in a peptide.

2.3. Training and validation

For the construction of the prediction models, the Random Forest algorithm (RF) was evaluated. The training of the models was carried out in the Python 3.6 programming language. The Anaconda 3 package (available at https://www.anaconda.com) was used to run the libraries: 'sklearn.ensemble', 'RandomForestClassifier', 'pandas', 'sklearn.externals', 'joblib' and 'score'. The 'score' function (accuracy) was implemented to choose models with scores > 0.95 as the cut-off for posterior validations.

The score function measures the accuracy of probabilistic predictions and ranges from 0 to 1. For model validations the following equations were used:

$$Sensitivity\ (TPR) = TP/(TP + FN)$$

$$Specificity\ (SPC) = TN/(TN + FP)$$

$$Accuracy\ (ACC) = TP + TN/(TP + FP + FN + TN)$$

where TP represents the true positives; TN the true negatives; FP the false positives and FN the false negatives. For the validation of the method, in addition to the equations mentioned above, the correlation coefficient of Matthews (MCC) was calculated:

$$MCC = (TP)(TN) - (FP)(FN) \Big/ \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

MCC is used to evaluate the performance of the predictor. Its value ranges from −1 to 1 and a larger MCC means a better prediction [22].

2.4. Software development

For the development of our application, we used the programming language Python 3.6 and the WinPython software which is a free open-source portable distribution of the Python programming language. AntiVPP 1.0 has a friendly interface that, in addition to having the ability to discriminate antiviral and non-antiviral peptides, can also be used to calculate different physical-chemical characteristics of the peptides. The software as well as the instructions to run it is available at https://github.com/bio-coding/AntiVPP.1.0.

3. Results

3.1. Training and validation

During the training with the data set T544p + 544n* we obtained several prediction models based on RF with scores > 0.95, each of these models were subjected to validation with the use of the independent data set V60p + 60n*. After evaluating each of the models obtained on the validation data, we selected a model with the best balance in the performance measures: TPR = 0.87, SPC = 0.97, ACC = 0.93 and MCC = 0.87. This model presented a score = 0.993 during the training phase.

Previously, we had performed an analysis using the Support vector machine (SVM), Artificial neural network (ANN) and k-nearest neighbor (kNN) algorithms in the prediction of antiviral peptides, observing a better balance in the performance measures obtained with the RF algorithm (Table 1).

3.2. Software development

Our software was developed with the programming language Python 3.6. AntiVPP 1.0 is an application with a simple and intuitive interface, making it ideal for researchers who are involved in the search and design of AVPs and they lack knowledge about the field of machine learning (Fig. 2). AntiVPP 1.0 returns two types of predictions: 'True' for positive cases and 'False' for negative cases. In addition, the software performs the computation of several peptide features, which are the

Table 1
Prediction models of antiviral peptides obtained by different algorithms on the validation dataset (V60p + 60n*).

Algorithm	Performance measurements			
	TPR	SPC	ACC	MCC
RF	0.87	0.97	0.93	0.87
SVM	0.85	0.93	0.79	0.84
ANN	0.87	0.95	0.90	0.85
kNN	0.83	0.91	0.90	0.81

TPR: sensitivity, **SPC:** specificity, **ACC:** accuracy, **MCC:** correlation coefficient of Matthews, **RF:** Random Forest, **SVM:** Support vector machine, **ANN:** Artificial neural network, **kNN:** k-nearest neighbor.

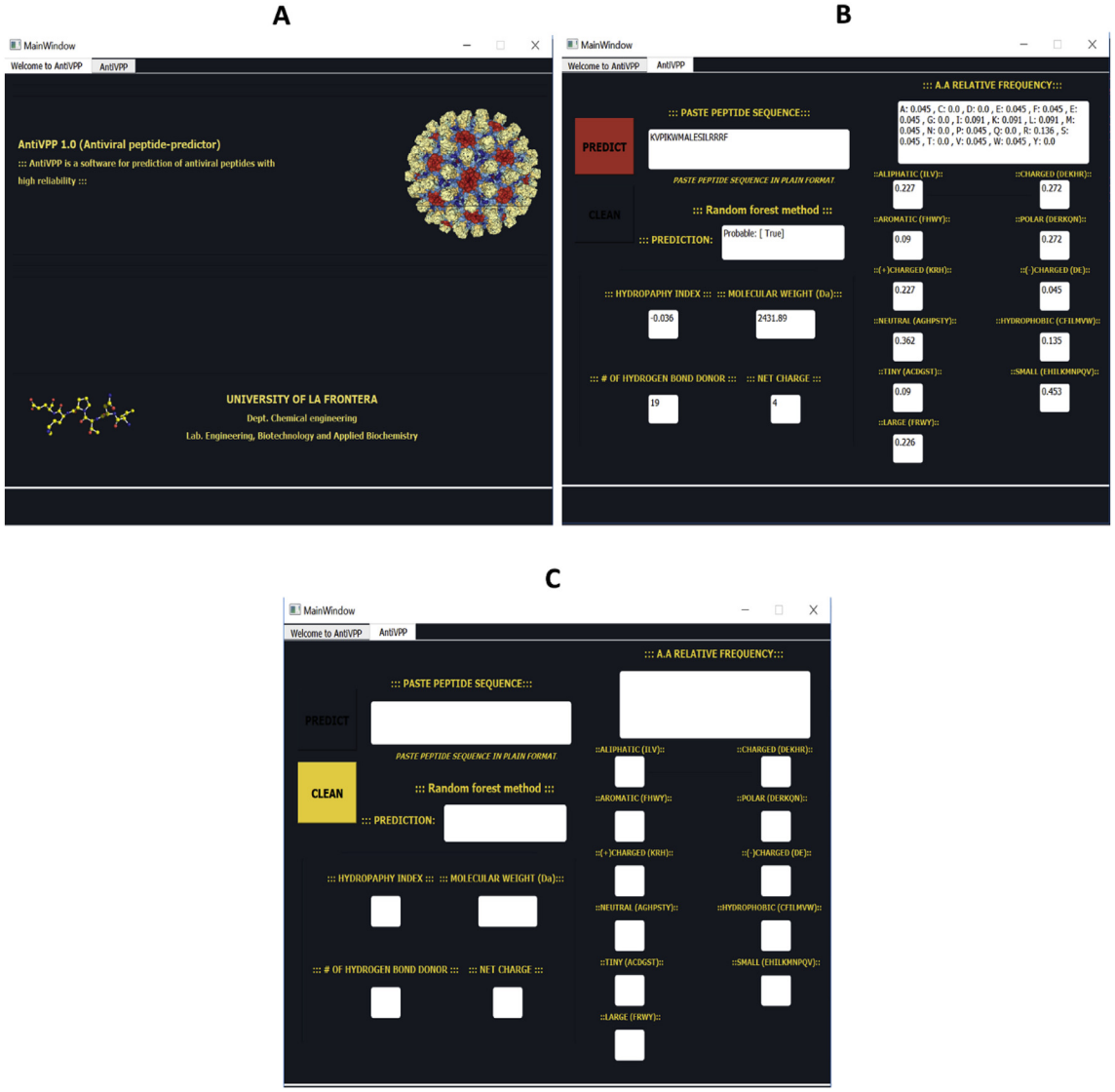

Fig. 2. Front of AntiVPP 1.0 (a). Button (PREDICT) for prediction of peptides in antiviral ['True'] or non-antiviral ['False'] (b). Button (CLEAN) to reset all the fields (c).

characteristics used for this program in AVPs classifications.

4. Discussion

Viral infections are one of the most important risks to consider for global health [23,24]. Over the last 50 years, extensive efforts have been dedicated to the development of antiviral drugs and great success has been accomplished for some viruses. Nevertheless, there are other viral infections such as epidemic influenza, which continue to spread worldwide and new threats of viruses, as well as drug-resistant viruses, are continuously emerging [23]. Peptide-based drugs have been of great interest to the scientific community from the past decade to the present, given that the modern pharmaceutical industry has come to appreciate the role of these molecules in addressing unmet medical needs. All this is because the peptides can be an excellent complement or even a more suitable alternative to small molecules and biological therapeutics [25]. Regardless of the potential of AVPs, there is a considerable lack of algorithms for AVPs prediction compared to other areas such as the investigation of antimicrobial peptides.

To date, the algorithm based on RF for the prediction of AVPs has been the one that has shown a better performance in the prediction of these molecules as reported in the literature [11,16,17]. The comparison of the performance measures obtained in our study, using the different algorithms, supports the previous results on the robustness of RF for AVP predictions [16], as shown in Table 1.

In this study, we evaluated the RF algorithm using new combinations of chemical-physical characteristics of the AVPs, obtaining an excellent model with the following performance measures during the validation phase: TPR = 0.87, SPC = 0.97, ACC = 0.93, and MCC = 0.87. In addition, we also confirmed the need to include the relative frequency for the improvement of AVP predictions as previously reported [16]. A comparison among the existing methods for the prediction of AVPs shows that AntiVPP 1.0 has the highest SPC. Specificity is one of the most relevant measures in the construction of predictive models and is characterized by determining the proportion of positive cases (AVPs) correctly identified (Table 2) [26].

On the other hand, we report for the first time the number of hydrogen bond donors as another important characteristic to be considered in the development of future AVP prediction algorithms, due to its role improving the quality of performance measures during the testing of our prediction models. It has been studied that H-bond pairing has a great influence on ligand-binding affinity, improving the strength of ligand-receptor interactions [27]. For this reason hydrogen bonds have had an important role in the design and discovery of new

Table 2
Comparison of the existing programs for prediction of AVPs.

Programs	Performance measurements				Ref.
	TPR	SPC	ACC	MCC	
AntiVPP 1.0	0.87	0.97	0.93	0.87	*
AVPpred	0.93	0.92	0.93	0.85	[11]
Model	0.93	0.93	0.93	0.87	[16]
IC50Pred	Not reported	Not reported	Not reported	Not reported	[17]

TPR: sensitivity, **SPC:** specificity, **ACC:** accuracy, **MCC:** correlation coefficient of Matthews, **RF:** Random Forest, **SVM:** Support vector machine, **ANN:** Artificial neural network, **kNN:** k-nearest neighbor, ***:** current study.

peptide-based drugs [28]. This feature is addressed in our work in a novel way, since it had not been used previously for the prediction of antiviral peptides.

5. Conclusion

AntiVPP 1.0 is a fast, accurate and intuitive tool focused on prediction of antiviral peptides as alternatives to the current tools for this purpose. The hydrogen bond is an important feature to consider in future algorithms addressed to the design and discovery of future antiviral peptides. This software would be helpful for researchers working in the development of antiviral therapies based on peptides due to its high success rates and user-friendliness.

Conflicts of interest

There is no conflict of interest to declare.

Notes

AntiVPP 1.0 is protected by copyright. This software is free for academic users. For commercial purposes, please contact: jorge.farias@ufrontera.cl.

Acknowledgments

This work was supported by the projects: DI12-PEO1 (EXE12-0004) DIUFRO and DIUFRO DIE14-0001 of the Universidad de La Frontera, Chile.

References

[1] H. Badani, R.F. Garry, W.C. Wimley, Peptide entry inhibitors of enveloped viruses: the importance of interfacial hydrophobicity, Biochim. Biophys. Acta Biomembr. 1838 (2014) 2180–2197.
[2] E.C. Holmes, Evolutionary history and phylogeography of human viruses, Annu. Rev. Microbiol. 62 (2008) 307–328.
[3] E. Domingo, Mechanisms of viral emergence, Vet. Res. 41 (2010) 38.
[4] K. Mulder, L.A. Lima, V. Miranda, S.C. Dias, O.L. Franco, Current scenario of peptide-based drugs: the key roles of cationic antitumor and antiviral peptides, Front. Microbiol. 4 (2013) 321.
[5] J.R. Shartouny, J. Jacob, Mining the Tree of Life: Host Defense Peptides as Antiviral Therapeutics, Seminars in Cell & Developmental Biology, Elsevier, 2018.
[6] P. Vlieghe, V. Lisowski, J. Martinez, M. Khrestchatisky, Synthetic therapeutic peptides: science and market, Drug Discov. Today 15 (2010) 40–56.
[7] T. Matthews, M. Salgo, M. Greenberg, J. Chung, R. DeMasi, D. Bolognesi, Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes, Nat. Rev. Drug Discov. 3 (2004) 215.
[8] E. Real, J.-C. Rain, V. Battaglia, C. Jallet, P. Perrin, N. Tordo, P. Chrisment, J. D'Alayer, P. Legrain, Y. Jacob, Antiviral drug discovery strategy using combinatorial libraries of structurally constrained peptides, J. Virol. 78 (2004) 7410–7417.
[9] S. Portal-Núñez, C.J. González-Navarro, M. García-Delgado, J.L. Vizmanos, J.J. Lasarte, F. Borrás-Cuesta, Peptide inhibitors of hepatitis C virus NS3 protease, Antiviral Chem. Chemother. 14 (2003) 225–233.
[10] H. Zhao, J. Zhou, K. Zhang, H. Chu, D. Liu, V.K.-M. Poon, C.C.-S. Chan, H.-C. Leung, N. Fai, Y.-P. Lin, A novel peptide with potent and broad-spectrum antiviral activities against multiple respiratory viruses, Sci. Rep. 6 (2016) 22008.
[11] N. Thakur, A. Qureshi, M. Kumar, AVPpred: collection and prediction of highly effective antiviral peptides, Nucleic Acids Res. 40 (2012) W199–W204.
[12] G. Wang, X. Li, Z. Wang, APD3: the antimicrobial peptide database as a tool for research and education, Nucleic Acids Res. 44 (2015) D1087–D1093.
[13] F.H. Waghu, R.S. Barai, P. Gurung, S. Idicula-Thomas, CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides, Nucleic Acids Res. 44 (2015) D1094–D1097.
[14] A. Qureshi, N. Thakur, M. Kumar, HIPdb: a database of experimentally validated HIV inhibiting peptides, PLoS One 8 (2013) e54908.
[15] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, Sci. Rep. 7 (2017) 42362.
[16] K.Y. Chang, J.-R. Yang, Analysis and prediction of highly effective antiviral peptides based on random Forest, PLoS One 8 (2013) e70166.
[17] A. Qureshi, H. Tandon, M. Kumar, AVP-IC50Pred: multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50), Pept. Sci. 104 (2015) 753–763.
[18] P. Klein, M. Kanehisa, C. DeLisi, Prediction of protein function from sequence properties: discriminant analysis of a data base, Biochim. Biophys. Acta Protein Struct. Mol. Enzymol. 787 (1984) 221–226.
[19] J.L. FAUCHÈRE, M. Charton, L.B. Kier, A. Verloop, V. Pliska, Amino acid side chain parameters for correlation studies in biology and pharmacology, Int. J. Pept. Protein Res. 32 (1988) 269–278.
[20] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 27 (1999) 368–369.
[21] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[22] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, PLoS One 12 (2017) e0177678.
[23] N. Takizawa, M. Yamasaki, Current landscape and future prospects of antiviral drugs derived from microbial products, J. Antibiot. 71 (2018) 45.
[24] B. McCloskey, O. Dar, A. Zumla, D.L. Heymann, Emerging infectious diseases and pandemic potential: status quo and reducing risk of global spread, Lancet Infect. Dis. 14 (2014) 1001–1010.
[25] A. Henninot, J.C. Collins, J.M. Nuss, The current state of peptide drug discovery: back to the future? J. Med. Chem. 61 (2017) 1382–1414.
[26] D.G. Altman, J.M. Bland, Diagnostic tests. 1: sensitivity and specificity, BMJ Br. Med. J. 308 (1994) 1552.
[27] D. Chen, N. Oezguen, P. Urvil, C. Ferguson, S.M. Dann, T.C. Savidge, Regulation of protein-ligand binding affinity by hydrogen bond pairing, Sci. Adv. 2 (2016) e1501240.
[28] D.J. Craik, D.P. Fairlie, S. Liras, D. Price, The future of peptide-based drugs, Chem. Biol. Drug Des. 81 (2013) 136–147.