scientific reports

Check for updates

OPEN

Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance

Abu Sayed Chowdhury[1], Sarah M. Reehl[2], Kylene Kehn-Hall[3,4,5], Barney Bishop[6] & Bobbie-Jo M. Webb-Robertson[1]✉

The emergence of viral epidemics throughout the world is of concern due to the scarcity of available effective antiviral therapeutics. The discovery of new antiviral therapies is imperative to address this challenge, and antiviral peptides (AVPs) represent a valuable resource for the development of novel therapies to combat viral infection. We present a new machine learning model to distinguish AVPs from non-AVPs using the most informative features derived from the physicochemical and structural properties of their amino acid sequences. To focus on those features that are most likely to contribute to antiviral performance, we filter potential features based on their importance for classification. These feature selection analyses suggest that secondary structure is the most important peptide sequence feature for predicting AVPs. Our Feature-Informed Reduced Machine Learning for Antiviral Peptide Prediction (FIRM-AVP) approach achieves a higher accuracy than either the model with all features or current state-of-the-art single classifiers. Understanding the features that are associated with AVP activity is a core need to identify and design new AVPs in novel systems. The FIRM-AVP code and standalone software package are available at https://github.com/pmartR/FIRM-AVP with an accompanying web application at https://msc-viz.emsl.pnnl.gov/AVPR.

Zoonotic viruses such as Ebola virus, Zika virus, West Nile virus and recently severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) can cause life-threatening disease outbreaks due to their high genetic diversity, variety of routes for transmission, and ability to replicate efficiently and to persist in their hosts[1–4]. The control of viral disease continues to be a challenging task due to increased resistance to available antiviral therapies, which are limited, and the continual emergence of novel viral pathogens. Antiviral peptides (AVPs) are a subset of antimicrobial peptides and are a potential resource for the development of new potent therapeutics for preventing or treating viral infection. The ability of AVPs to target various aspects of the viral lifecycle, ranging from their attachment to host cells to their ability to impair viral replication within the cells has been the subject of multiple studies[5–13]. AVPs can be natural or synthetic, obtained by introducing chemical groups or non-natural amino acids into natural peptide sequences[4,13,14]. Considering AVPs in the design of new antiviral therapeutics is advantageous because it allows us to capitalize on their low molecular weight, low toxicity, high specificity and effectiveness, and minor side effects[15]. Machine learning is a powerful strategy for identifying AVPs by leveraging the ever-increasing data available in public databases, such as AVP Prediction (AVPpred)[16], Antimicrobial Peptide Database (APD3)[17], Collection of Antimicrobial Peptides (CAMPR3)[18] and HIV inhibitory peptides database (HIPdb)[19].

Researchers have previously developed machine learning models[16,20–25] for predicting AVPs. Thakur et al.[16] developed AVPpred, a web server for collecting and detecting highly effective AVPs. The authors used a support vector machine (SVM) to build two machine learning models based on amino acid composition (AAC) and physicochemical features. This was then extended to use a random forest (RF)-based model[20], which was able to outperform the SVM utilized in AVPpred. The RF models were constructed using AAC, physicochemical

[1]Biological Sciences Division, Pacific Northwest National Laboratory, J4-18, P.O. Box 999, Richland, WA 99354, USA. [2]Computing and Analytics Division, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99354, USA. [3]School of Systems Biology, George Mason University, Manassas, VA 20110, USA. [4]National Center for Biodefense and Infectious Diseases, George Mason University, Manassas, VA 20110, USA. [5]Department of Biomedical Sciences and Pathobiology, Virginia Tech, Blacksburg, VA 24061, USA. [6]Department of Chemistry and Biochemistry, George Mason University, Manassas, VA 20110, USA. ✉email: bj@pnnl.gov

properties, aggregation propensities of amino acids and secondary structure. Lissabet et al.[21] developed a portable software version of the RF method called AntiVPP 1.0 that gives improved prediction accuracy. Qureshi et al.[22] introduced a regression-based algorithm AVP-IC_{50}Pred to predict AVP half maximal inhibitory concentration (IC_{50}). Various peptide features such as AAC, binary profile, physicochemical properties, solvent accessibility were considered, and a number of machine learning techniques with individual and different combination of features were used to predict the IC_{50} value of the peptide sequences. Further, based on the assumption that AVPs have low sequence similarity the use of pseudo amino acid composition (PseAAC)[26] was introduced as AVP peptide features in the AdaBoost machine learning model[23]. In recent years ensemble-based methods have been introduced, such as Meta-iAVP[25] and PePred-Suite[24]. The Meta-iAVP approach uses machine learning to transform the feature space into a modified 6-dimensional predicted output vector, which then becomes the input data to the meta-classifier to predict the class of validation data set. PEPred-Suite is similar to Meta-iAVP where a RF is used as both the base and meta classifiers. Both Meta-iAVP and PEPred-Suite use these ensemble strategies to improve the AVP prediction accuracy.

The series of machine learning developments in AVP have to date focused on increasing the features that characterize a peptide and making minor modifications to the machine learning algorithm. They have not included feature reduction techniques that would determine the most relevant and non-redundant features from the initial set of input features. The performance of a machine learning model can rely heavily on using the most informative features, with the inclusion of non-informative features resulting in potential degradation in classifier performance. In the current study we identified the most important features by estimating Pearson's correlation coefficient and mean decrease of Gini index (MDGI) for all candidate features, which is a metric of feature importance based on the individual decision trees in a random forest model. The candidate features were generated from the physicochemical and secondary structure properties of a library of known AVP and non-AVP sequences. Subsequently, we applied a recursive feature elimination (RFE) algorithm in combination with the SVM to determine the order of importance of the different features. We evaluated multiple machine learning approaches, including SVM, RF and deep learning (DL) via multiple neural network architectures and hyperparameters, for training and testing purposes using our selected feature set. Our SVM-based method achieved the best test accuracy and Matthews correlation coefficient (MCC) values compared to the RF and DL approaches as well as outperformed AVPpred[16] and Chang et al.'s method[20]. We packaged the resulting approach into a software tool called Feature-Informed Reduced Machine Learning for Antiviral Peptide Prediction (FIRM-AVP).

Methods

Training and testing data. We used the same experimentally validated dataset reported in AVPpred[16] that has been used consistently since its introduction to evaluate AVP prediction models. It consists of a total of 1056 unique peptides. This set of peptides was distilled from a starting collection of 1245 peptides that were reduced to remove peptides with too high of similarity. Out of them, 604 sequences are highly effective (positive samples), and 452 sequences are minimally or non-effective AVPs (negative samples). These datasets were used for training and validating the machine learning model. To construct the training and independent test sets to benchmark our results with existing SVM and RF-based models we followed the same process as described previously[16,20]. This yields 544 and 407 positive and negative samples in the training dataset, respectively, and the validation/independent test set consisted of 60 and 45 positive and negative samples, respectively as defined by prior publications to assure accurate comparison. This validation set has similar overall viral diversity as the training set. On the AVPpred server there are additional peptides for the negative samples set, 544 in training set and 60 in the independent test set, however; these peptides have not been confirmed experimentally and thus are not included here.

Feature generation. We combined several sets of features based on the peptide sequences: a 20D feature vector for AAC expressed as the percentage representation of a particular amino acid in a peptide; a 400D feature set was generated based on the dipeptide composition (DC) which represents the fraction of dipeptides within a peptide sequence; and the PseAAC and amphiphilic pseudo amino acid composition (APseAAC) proposed by Chou[26,27] to incorporate sequence-order information. The dimension of the PseAAC feature vector is $20 + \boldsymbol{\lambda} \times \boldsymbol{\omega}$ where $\boldsymbol{\lambda}$ is the discrete correlation factor and $\boldsymbol{\omega}$ is the weight factor of the sequence information. In our case, we set $\boldsymbol{\lambda} = 5$ and $\boldsymbol{\omega} = 0.05$ by considering the minimum length of our collected AVP and non-AVP sequences. So, in the 25D PseAAC feature vector, the first 20 features are the traditional AAC and the other components are the rank-different correlation factors that represents the sequence-order information. We produced a $20 + 2\boldsymbol{\lambda}$ i.e., 30D, APseAAC feature vector where the first 20 features are the basic AAC and the remaining components indicate the correlation factor for the physicochemical properties of peptides. We also utilized the composition, transition, and distribution (CTD) model[28–31] to generate feature vectors for 8 physicochemical properties; hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structure, solvent accessibility and surface tension of peptide sequences. In the CTD model, amino acids are classified into three classes based on their physicochemical properties. For composition, we obtained 3D feature vector that give the fraction of each encoded class in a peptide sequence. A transition feature vector of 3D gives the transition of one class followed by another class and vice versa. We also obtained a 15D feature vector for distribution that indicates the percent distribution (i.e., 1%, 25%, 50%, 75% and 100%) of each class in a peptide sequence. As we have 8 physicochemical properties, the CTD model gives a $(3+3+15)\times 8 = 168$D feature vector. Finally, we retrieved features from the secondary structure of peptide sequences. A total of six features were extracted from the location information, spatially consecutive states and segment sequences of the three main types of secondary structure; $\boldsymbol{\alpha}$-helix, $\boldsymbol{\beta}$-strand and $\boldsymbol{\gamma}$-coil. The details of feature extraction from the CTD model and secondary structure information of amino acid sequences were explained in our previous works[32–34]. In summary, we generated 649

Peptide feature	Feature dimension
Amino acid composition	20
Dipeptide composition	400
Pseudo-amino acid composition	25
Amphiphilic pseudo-amino acid composition	30
Composition/transition/distribution	168
Secondary structure sequence	6

Table 1. List of 649 peptide features.

peptide sequence-based features listed in Table 1 using the R programming language (*ver* 4.0.0)[35]. We utilized the *protr* (*ver*. 1.6-2)[30] and *DECIPHER* (*ver*. 2.14.0)[36] packages to extract features from peptide sequences.

The DC feature vector (dipep_1, dipep_2, …, dipep_400) are the dipeptide composition (Supplementary Table S1) of the amino acids in order A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V. The PseAAC and APseAAC are the feature vectors (pseudo_1, pseudo_2, …, pseudo_25) and (amphipseudo_1, amphipseudo_2, …, amphipseudo_30), respectively. The composition feature vector (comp_1, comp_2, …, comp_24) and transition feature vector (tran_1, tran_2, …, tran_24) are the composition and transition values in the order-physicochemical property 1 (group 1), physicochemical property 1 (group 2), physicochemical property 1 (group 3) and so on. In the distribution feature vector (dist_1, dist_2, …, dist_120), the first 15D features are the group1, group2 and group3 distribution values for the first physiochemical property and so on. The physicochemical properties and their groups are listed as supplementary Table S2. Finally, in the 6D secondary structure feature vector, ss_1, ss_2 and ss_3 are the location-oriented features for the $\boldsymbol{\alpha}$-helix, $\boldsymbol{\beta}$-strand and $\boldsymbol{\gamma}$-coil, respectively. The other three features ss_4, ss_5 and ss_6 gives the normalized maximum spatial consecutive $\boldsymbol{\alpha}$-helix and $\boldsymbol{\beta}$-strand in the secondary structure sequence, and occurrences of segmented sequences "$\boldsymbol{\beta}$-strand $\boldsymbol{\alpha}$-helix $\boldsymbol{\beta}$-strand" after ignoring $\boldsymbol{\gamma}$-coil states from the secondary structure.

Machine learning models. We utilized three machine learning approaches to train the AVP prediction model, traditional SVM and RF methods, as well as DL via multiple architectures and hyperparameters using the machine learning library, *caret* (*ver.* 6.0-86)[37]. For the DL, variations on the Multi-layer Perceptron were the most successful. These binary classification models were then used to classify the test set of peptides. Note that we tuned the SVM and RF models with the training dataset and used the best models for prediction. The SVM model was tuned using the radial basis function kernel with cost values of 4, 8, 16, 32, 64, and 128. The RF model was tuned with *ntree* values of 50, 100, 200, 300, 400 and 500 and *mtry* values of 2, 4, 8, 16, and 32. The final SVM model used a *cost* value of 8, and RF model was with *ntree* = 100 and *mtry* = 32, which was chosen as best models for the selected feature on the training data. We utilized the *e1071* (*ver.* 1.7-3)[38] package to tune the models.

Feature selection. The 649 features may contain redundant and information irrelavent to the classification of AVPs. To reduce the dimensions of the features we calculated the Pearson's correlation coefficient [using Eq. (1)] between two feature vectors x and y across all of the peptides to observe the linear correlation between features. Here E, μ and σ are the expectation, mean and standard deviation values, respectively.

$$\rho = \frac{E\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x \sigma_y}. \tag{1}$$

If the absolute value of the correlation between two features is greater than a threshold value, one of the two features were removed randomly from further consideration. We considered a range of correlation threshold from 0.7 to 0.95 in increments of 0.05. A correlation threshold was selected to optimize the Area Under a Receiver Operating Characteristic Curve (AUC) associated with the feature selection, which set the parameter to 0.85 and reduced the dataset to 568 features. We utilized the R *stats* package (*ver*. 3.6.2) to compute the Pearson correlation values between features.

As a next step, we computed mean decrease of Gini index (MDGI) using an RF model for the remaining features. We can find the feature importance using MDGI to measure the contribution of each feature to the homogeneity of the nodes and leaves in the RF model[39]. A node is considered as more pure in the RF model if the Gini index is closer to 0. The Gini index is calculated using Eq. (2) where we subtract the sum of the squared probabilities of each of the two classes from 1.

$$Gini = 1 - \sum_{i=1}^{2} P_i^2. \tag{2}$$

So, the Gini index values of 0 and 1 indicate completely homogeneous data and completely heterogeneous data, respectively. To find the feature importance, whenever a feature is used to divide data at a node, we calculated the Gini index at the root node and at both the leaves. The difference in the Gini index of splitting root node and weighted Gini index of the child nodes was estimated to find the fall of Gini index values in a decision tree of the RF model[20]. For each feature, MDGI is the average value of all the decrease of Gini index over all the decision trees created in the RF model and higher MDGI value indicate elevated feature importance. Based on

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
FIRM-AVP (SVM)	93.3	91.1	92.4	0.84
FIRM-AVP (RF)	95.0	82.2	89.5	0.79
FIRM-AVP (DL)	91.7	80.0	86.7	0.73
AVP-649D (SVM)	95.0	82.2	89.5	0.79
AVP-649D (RF)	90.0	82.2	86.7	0.73
AVPcompo	83.3	88.9	85.7	0.72
AVPphysico	88.3	82.2	85.7	0.71
RFcompo + structure + agg	91.7	86.7	89.5	0.79
Meta-iAVP	95.2	96.7	93.2	0.90

Table 2. Performance comparison of our models with existing models on independent validation data.

the MDGI we down-selected to 169 features with positive MDGI. The *randomForest* (*ver*. 4.6-14)[40] package was used to estimate the MDGI values of the features.

Recursive feature elimination. Following reduction of the number of features based on Pearson's correlation coefficient and MDGI values, we applied the RFE technique[41] to the machine learning models using the training data for the reduced feature set to order the features by importance. RFE evaluates the training performance of a machine learning model for a feature set and gives the ranking of the features. We considered 10-fold cross validation with 5 repeats to evaluate the training performance of the machine learning models. We utilized *caret* (*ver*. 6.0-86)[37] to implement the RFE algorithm.

Performance measurement. We utilize the area under the receiver operating characteristic (ROC) curve (AUC) values to measure the training performance of the models via RFE for the reduced feature set. ROC curves use a combination of the true positive rate and false positive rate to provide a summary of the prediction capability of a machine learning model where a perfect classifier has an AUC of 1.0 and a random binary classifier will have an AUC of 0.5. We report the final test performance of our classifiers using the same metrics as previously reported for other AVP prediction algorithms, which include sensitivity, specificity, accuracy and MCC values [Eqs. (3–6)], where TP, TN, FP, and FN are true positives (positives accurately classified), true negatives (negatives accurately classified), false positives (negatives classified as positives), and false negatives (positives classified as negatives), respectively. The MCC value is used to evaluate the efficacy of a classifier as the number of positive and negative examples in the datasets is imbalanced and the range of this value is [− 1, 1]. Higher MCC value indicates better prediction.

$$Sensitivity = \frac{TP}{TP + FN}, \tag{3}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{6}$$

Data availability. All experimental data are available at https://github.com/pmartR/FIRM-AVP.

Results

AVP prediction performance. The performance of the FIRM-AVP SVM, RF and DL models were compared based on the standard metrics of sensitivity, specificity, accuracy, and MCC for their performance on the validation/independent dataset where a positive AVP peptide is defined as a probability of greater than 0.5 and a negative AVP as less than or equal to 0.5. Evaluating overall accuracy, we observe that the SVM and RF models have very high AUC values, 0.962 and 0.958, respectively. Table 2 details the results of the models for the 169 features based on our feature reduction. The SVM model achieved 92.4% accuracy and 0.84 MCC, which is better than the RF model. Both the SVM and RF machine learning approaches yield a posterior probability that represent the probability that a peptide is AVP given the data represented as the 169 features, or likewise the probability that a peptide is non-AVP. We evaluated the probabilities of the 60 positive AVPs for the SVM versus the RF and found that on average the strength of the prediction based on the probability for the SVM was larger than the RF by ~ 0.02 (paired t-test p-value ~ 0.14). Thus, there is marginal evidence that the SVM yields

Feature rank	Features for SVM	Features for RF
1	**Location oriented feature for α-helix**	Distribution (25% residues) feature for positive charge (group 1)
2	**PseAAC feature for leucine (L) amino acid**	Composition feature for intermediate solvent accessibility (group 3)
3	PseAAC feature for *isoleucine* (I) amino acid	**PseAAC feature for leucine (L) amino acid**
4	**PseAAC feature for lysine (K) amino acid**	**Location oriented feature for α-helix**
5	Composition feature for neutral hydrophobicity (group 2)	**PseAAC feature for lysine (K) amino acid**

Table 3. Top-5 features obtained in SVM and RF methods from RFE analysis. Common features are highlighted in bold.

a more confident identification, but it is not statistically significant based on this data at a p-value threshold of 0.05. However, when evaluating the negative class there is a significant improvement gained by the SVM. The average non-AVP peptide is generally correctly classified with a larger probability by ~ 0.09 (p-value ~ 5E−5). This difference in strength of classification of the non-AVP class is what is largely driving the reduction in false positives for the SVM, which is observed in the specificity values reported in Table 2. The DL approaches were likely sub-optimal because while multiple nonlinearities exist in these data, the training examples are too few to both describe the nonlinearities and adequately generalize to new data. Evidence of such a conclusion is apparent in discrepancies between training and testing loss, even in the presence of regularization. Future work of importance is to grow and create more variety in the AVP benchmark dataset, which has not been updated in 8 years, which would aid in the application of more recent machine learning approaches.

For the independent test set, we then compared the performance of our FIRM-AVP SVM model with no feature reduction (AVP-649D), as well as with the AVPpred[16] and the Chang et al.'s RF approaches (RFcompo + structure + agg)[20] (Table 2). There is a clear increase in accuracy based on the reduced feature set from the full 649 features, for example our best performing SVM model increased the MCC from 0.79 to 0.84 by reducing to the highest importance features. In terms of prior analyses, the AVPcompo and AVPphysico are the models of AVPpred based on AAC and physicochemical features, respectively whereas RFcompo + structure + agg is the Chang et al.'s RF method that uses both (AAC), secondary structure and aggregation features. Chang et al.'s RF method outperforms AVPpred with an accuracy of 89.5% and 0.79 MCC value. However, our FIRM-AVP SVM models that is built on an optimized feature set performed better than either of these two methods in terms of accuracy and MCC and the FIRM-AVP RF model was similar to that of prior models. The most accurate model is MetaiAVP[25], which is based on an ensemble of machine learning algorithms. This however comes with a challenge in interpretation and gaining insight into the features that are driving antiviral activity as was the goal with FIRM-AVP. The same validation set run on each of the 6 machine learning algorithms separately have MCC values that range from 0.34 to 0.73, well below the FIRM-AVP using a single classifier on the optimized feature set.

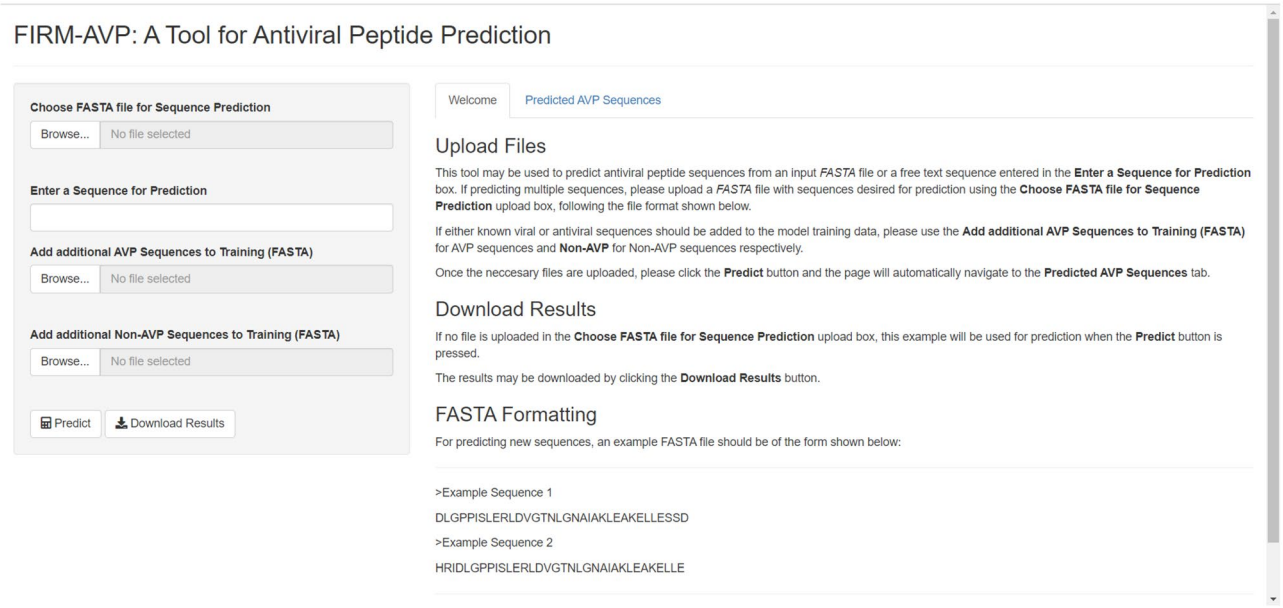
Recursive feature rankings. We performed RFE operations on the SVM model with the training data using 169 features from the initial feature selection with repetition measure the training performance of the SVM (in terms of AUC) via RFE algorithm. Note that the AUC values gradually decreased as features were removed from the model as depicted in Supplementary Fig. S1, and we obtained the highest AUC values of 0.89 and 0.92 for the SVM and RF models, respectively, by including all 169 features. This indicates that we do not need further feature reduction, and thus we utilize the RFE results to sort the importance of the features. Table 3 lists the top-5 features found after RFE analysis. Both secondary structure, composition and PseAAC features are in the top-5 features for both machine learning models. Peptide secondary structure features are identified as top ranked features in SVM and RF methods, respectively. All rankings of the selected features for both SVM and RF models are listed in Supplementary Table S3.

Software tool and user's manual. We developed the standalone software tool, FIRM-AVP based on the SVM algorithm. The open source software are available at https://github.com/pmartR/FIRM-AVP. Additionally, a web-based version of the software is available at https://msc-viz.emsl.gov/AVPR/ . To use the web application the users need to provide either a single peptide sequence or a FASTA file of peptide sequences to be analyzed and predictions will be returned that include the probability that a peptide sequence is antiviral (Fig. 1). As previously mentioned, a current limitation in AVP prediction is the scale of the data available on which to build predictive models. To make the software more useful for those working on improving the algorithm via collecting additional training data, the software provides the user an option to add new known AVP and non-AVP sequences to retrain the machine learning model. A simple page refresh will reset the model. The graphical user interface and options the web application provides are shown in Fig. 1. The feature generation and selection components of the software were implemented using R. The graphical user interface design and implementation were created using the R web application framework *shiny* (*ver*. 1.4.0.2)[42].

Discussion

Identifying potential AVPs is of great importance for the discovery of new drugs to treat viral infection. In this work, we introduced a machine learning model for predicting AVPs using a core set of 169 features identified via correlation and machine learning analyses. Our SVM and RF models were developed based on the features

(A)



(B)

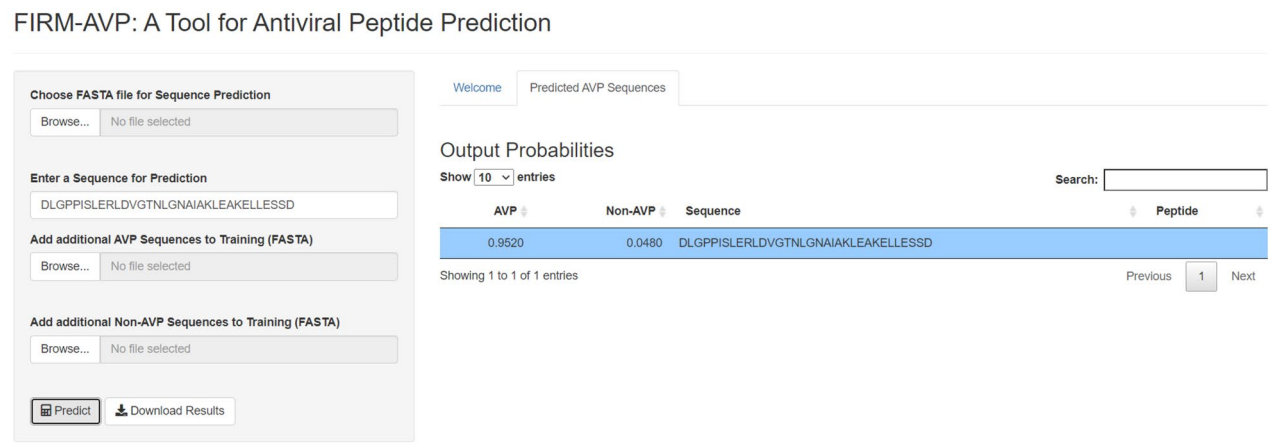


Figure 1. Online FIRM-AVP software interface (https://msc-viz.emsl.gov/AVPR/). Where (**A**) is the starting page that allows users to either paste in a single peptide sequence or upload a FASTA file containing a collection of peptide sequences. Example sequences and files are given. (**B**) The probability of AVP versus non-AVP is returned for each sequence based on the pasted peptide sequence or the uploaded FASTA file.

generated from the AAC, DC, PseAAC, APseAAC, CTD, and predicted secondary structure properties of peptide sequences. To verify the effectiveness of our best feature sets, we tested the performance of our models using an independent dataset that included the same validation/independent as prior methods[16,20]. We achieved higher accuracies and MCC values relative to single classifier models that did not include feature reduction, as well as existing published models, demonstrating the effectiveness of the feature selection approach. The software tool FIRM-AVP based on our approach is publicly available for user with flexible options to not only make predictions, but to update the underlying prediction model. The need for more training data was a limiting factor to the DL approach, which had lower overall accuracy than the SVM and RF approaches.

We evaluated multivariate feature importance using our selected feature sets via RFE. Secondary structure and distribution features were identified as top ranked features in our SVM and RF models, respectively. Location oriented features for α-helix conformation and distributional features associated with positive charge as the most important features of the machine learning models. The PSeAAC feature for leucine and lysine amino acids were also important in distinguishing AVP and non-AVP sequences. The location oriented feature for α-helix and PSeAAC features for leucine and lysine amino acids support the abundance of the α-helix structure, and leucine and lysine residues in AVPs that were claimed in the RF-based method[20] and HIPdb[19]. The observed significance of α-helical structure is consistent with the fact that many known antimicrobial peptides exhibit

varied degrees of helical conformation and spatial partitioning of cationic and hydrophobic residues[43]. Here, both the SVM and RF approaches establish helix distributional features that are associated with antiviral peptides[44,45]. How these properties factor in peptide antiviral activity is not clear, however they are known to contribute to their interactions with cell membranes.

The discovery of new antiviral therapies is imperative to address the challenge of new viral epidemics and AVPs can be a valuable resource for the development of novel therapies to combat viral infection. One of the core needs is not only improving the accuracy of AVP prediction models, but also building explainable models that can aid in understanding the fundamental multivariate properties that are associated with anti-viral activity. This is a necessary step in the design of AVP design for novel viral systems.

Received: 17 July 2020; Accepted: 20 October 2020
Published online: 06 November 2020

References

1. Domingo, E. Mechanisms of viral emergence. *Vet. Res.* **41**, 38 (2010).
2. Nichol, S. T., Arikawa, J. & Kawaoka, Y. Emerging viral diseases. *Proc. Natl. Acad. Sci.* **97**, 12411–12412 (2000).
3. Phan, T. Genetic diversity and evolution of SARS-CoV-2. *Infect. Genet. Evol.* **81**, 104260 (2020).
4. Qureshi, A., Thakur, N., Tandon, H. & Kumar, M. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res.* **42**, D1147–D1153 (2014).
5. Gleenberg, I. O., Avidan, O., Goldgur, Y., Herschhorn, A. & Hizi, A. Peptides derived from the reverse transcriptase of human immunodeficiency virus type 1 as novel inhibitors of the viral integrase. *J. Biol. Chem.* **280**, 21987–21996 (2005).
6. Gleenberg, I. O., Herschhorn, A. & Hizi, A. Inhibition of the activities of reverse transcriptase and integrase of human immunodeficiency virus type-1 by peptides derived from the homologous viral protein R (Vpr). *J. Mol. Biol.* **369**, 1230–1243 (2007).
7. Littler, E. & Oberg, B. Achievements and challenges in antiviral drug discovery. *Antiviral Chem. Chemother.* **16**, 155–168 (2005).
8. Louis, J. M., Dyda, F., Nashed, N. T., Kimmel, A. R. & Davies, D. R. Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. *Biochemistry* **37**, 2105–2110 (1998).
9. Pang, W., Tam, S.-C. & Zheng, Y.-T. Current peptide HIV type-1 fusion inhibitors. *Antiviral Chem. Chemother.* **20**, 1–18 (2009).
10. Rausch, D. *et al.* Peptides derived from the CDR3-homologous domain of the CD4 molecule are specific inhibitors of HIV-1 and SIV infection, virus-induced cell fusion, and postinfection viral transmission in vitro. Implications for the design of small peptide anti-HIV therapeutic agents. *Ann. N. Y. Acad. Sci.* **616**, 125–148 (1990).
11. Reusser, P. Antiviral therapy: Current options and challenges. *Schweizerische medizinische Wochenschrift* **130**, 101–112 (2000).
12. Prusoff, W. H., Lin, T., August, E. M., Wood, T. G. & Marongiu, M. E. Approaches to antiviral drug development. *Yale J. Biol. Med.* **62**, 215 (1989).
13. Qureshi, A., Kaur, G. & Kumar, M. AVC pred: An integrated web server for prediction and design of antiviral compounds. *Chem. Biol. Drug Des.* **89**, 74–83 (2017).
14. Boas, L. C. P. V., Campos, M. L., Berlanda, R. L. A., de Carvalho Neves, N. & Franco, O. L. Antiviral peptides as promising therapeutic drugs. *Cell. Mol. Life Sci.* **76**, 1–18 (2019).
15. Castel, G., Chtéoui, M., Heyd, B. & Tordo, N. Phage display of combinatorial peptide libraries: Application to antiviral research. *Molecules* **16**, 3499–3518 (2011).
16. Thakur, N., Qureshi, A. & Kumar, M. AVPpred: Collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res.* **40**, W199–W204 (2012).
17. Wang, G., Li, X. & Wang, Z. APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–D1093 (2016).
18. Waghu, F. H., Barai, R. S., Gurung, P. & Idicula-Thomas, S. CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.* **44**, D1094–D1097 (2016).
19. Qureshi, A., Thakur, N. & Kumar, M. HIPdb: A database of experimentally validated HIV inhibiting peptides. *PLoS ONE* **8**, e54908 (2013).
20. Chang, K. Y. & Yang, J.-R. Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS ONE* **8**, e70166 (2013).
21. Lissabet, J. F. B., Belén, L. H. & Farias, J. G. AntiVPP 1.0: A portable tool for prediction of antiviral peptides. *Comput. Biol. Med.* **107**, 127–130 (2019).
22. Qureshi, A., Tandon, H. & Kumar, M. AVP-IC50Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50). *Pept. Sci.* **104**, 753–763 (2015).
23. Zare, M., Mohabatkar, H., Faramarzi, F. K., Beigi, M. M. & Behbahani, M. Using Chou's pseudo amino acid composition and machine learning method to predict the antiviral peptides. *Open Bioinform. J.* **9**, 13–19 (2015).
24. Wei, L., Zhou, C., Su, R. & Zou, Q. PEPred-Suite: Improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics* **35**, 4272–4280 (2019).
25. Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V. & Shoombuatong, W. Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int. J. Mol. Sci.* **20**, 5743 (2019).
26. Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Bioinform.* **43**, 246–255 (2001).
27. Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **21**, 10–19 (2005).
28. Dubchak, I., Muchnik, I., Holbrook, S. R. & Kim, S.-H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci.* **92**, 8700–8704 (1995).
29. Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. & Kim, S. H. Recognition of a protein fold in the context of the SCOP classification. *Proteins Struct. Funct. Bioinform.* **35**, 401–407 (1999).
30. Xiao, N., Cao, D.-S., Zhu, M.-F. & Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **31**, 1857–1859 (2015).
31. Li, Z.-R. *et al.* PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **34**, W32–W37 (2006).
32. Chowdhury, A. S., Call, D. R. & Broschat, S. L. Antimicrobial resistance prediction for Gram-negative Bacteria via Game theory-Based feature evaluation. *Sci. Rep.* **9**, 1–9 (2019).
33. Chowdhury, A. S., Khaledian, E. & Broschat, S. L. Capreomycin resistance prediction in two species of Mycobacterium using a stacked ensemble method. *J. Appl. Microbiol.* **127**, 1656–1664 (2019).
34. Chowdhury, A. S., Call, D. R. & Broschat, S. L. PARGT: A software tool for predicting antimicrobial resistance in bacteria. *Sci. Rep.* **10**, 1–7 (2020).

35. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2020).
36. Wright, E. S. Using DECIPHER v2.0 to analyze big biological sequence data in R. *R J.* **8**, 352–359 (2016).
37. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
38. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2019).
39. Calle, M. L. & Urrea, V. Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* **12**, 86–89 (2011).
40. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
41. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
42. shiny: Web Application Framework for R (2020).
43. Huang, Y., Huang, J. & Chen, Y. Alpha-helical cationic antimicrobial peptides: Relationships of structure and function. *Protein Cell* **1**, 143–152. https://doi.org/10.1007/s13238-010-0004-3 (2010).
44. Tossi, A., Sandri, L. & Giangaspero, A. Amphipathic, alpha-helical antimicrobial peptides. *Biopolymers* **55**, 4–30. https://doi.org/10.1002/1097-0282(2000)55:1%3c4::AID-BIP30%3e3.0.CO;2-M (2000).
45. Zelezetsky, I. & Tossi, A. Alpha-helical antimicrobial peptides–using a sequence template to guide structure-activity relationship studies. *Biochim. Biophys. Acta* **1758**, 1436–1449. https://doi.org/10.1016/j.bbamem.2006.03.021 (2006).

Acknowledgements

This work was supported by the U.S. Army Medical Research Acquisition Activity, through the Accelerating Innovation in Military Medicine program under Award No. W81XWH-18-1-0801. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense or the U.S. Army. We are grateful to the St. Augustine Alligator Farm Zoological Park for their collaboration on this project. Computational work was completed at Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle Memorial Institute for the Department of Energy under contract DEAC05-76RLO1830.

Author contributions

A.S.C., S.M.R., and B.J.W.R developed and evaluated the machine learning algorithms. B.B. and K.K.H. provided feature evaluation and interpretation. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-76161-8.

Correspondence and requests for materials should be addressed to B.-J.M.W.-R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.