EnAVP-Class: An Ensemble Deep Learning Model for Characterizing Antiviral Activity in Peptides

Caio Fontes de Castro*[1,2]; Fernanda Midori Abukawa[1]; Milton Y. Nishiyama-Jr*[1]

[1]LETA- Laboratory of Applied Toxinology, Butantan Institute, Brazil. [2]Institute of Mathematics and Statistics, University of São Paulo, Brazil.
*E-mail corresponding author: caiofcastro@usp.br; milton.nishiyama@butantan.gov.br

INTRODUCTION

Bioactive Peptides (AMPs) have been the targets of growing interest due to their presence in virtually all living beings, and potential for the development of new drugs with a lower probability of resistance developing. In particular Peptides with Antiviral activity (AVPs) have been studied as alternatives to traditional antiviral drugs, which commonly have severe adverse side effects.

Venomous organisms frequently utilize AMPs in their venoms and toxins, but most of the compounds present in these mixtures are not fully understood or characterized. In particular the venom from arachnids is believed to be a potential source of undiscovered bioactive molecules.

Computational methods have been used extensively in the identification of bioactive peptides of many functional types. Models based in Neural Networks specially, have achieved better results.

In this project we propose the development of a Two-Phase classifier, in which the first phase is based on an Ensemble model of a Deep Neural Network and a Random Forest module, capable of predicting the antiviral activity of unknown peptides based on Proteomic and Transcriptomic data. The second phase is a ML model which in turn classifieds the AVPs by their antiviral activity kind.

This tool could assist researchers in the identification of future antiviral compounds.

METHODOLOGY

We gathered experimentally validated AMP sequences from 7 open databases, including APD3, CAMP, DBAASp etc, totalling **15687 AMPs** of which **2812 are AVPs**. For the negative dataset we gathered proteins from the reviewed Uniprot Database, and selected proteins without any function related to the activities of interest (in this case antimicrobial or antiviral). These proteins were then digested *in silico* with Trypsin to generate the putative non-AMP peptides. Both sets of peptides went through a homology reduction using *cd-hit* tool, and then we selected peptides with no Bioactivity in the same length distribution as the AVPs to form the negative classes. We then selected the pepdtides with known targets and split then in 3 classes: “Membrane”, “Replication” and “Viral Assembly”. The whole process can be seen in Figure (1). The model architecture can be seem in Figure (2).

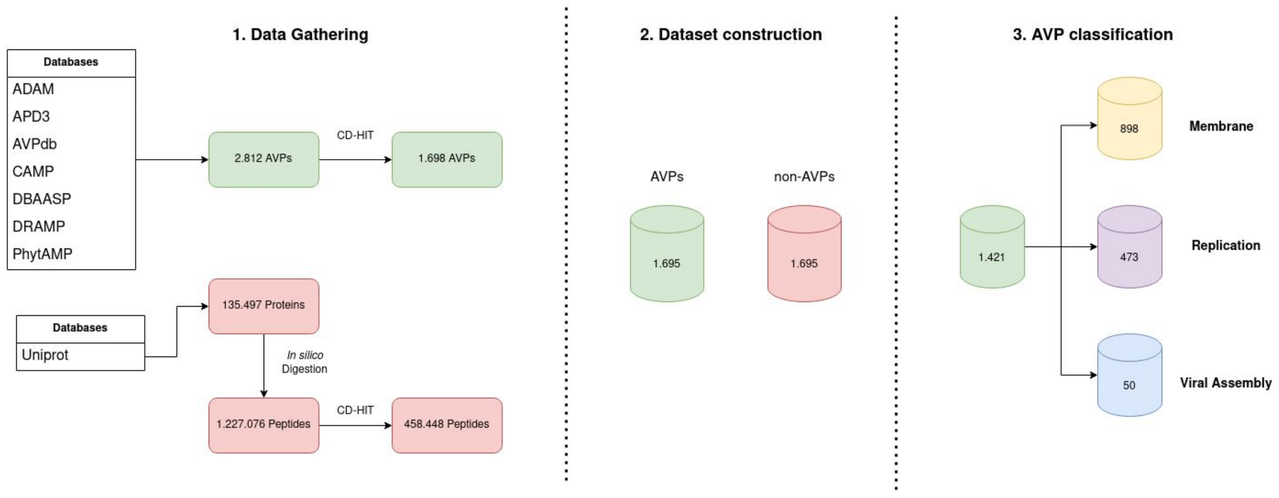

Fig. 1 - Construction of the Datasets. 1: Initial gathering of AMPs and AVPs from open repositories and first pre-processing steps. 2: Initial datasets. Negative sequences were selected in the same length distribution and positive ones. 3: Utilizing literature information to characterize the action mechanism of the peptides.

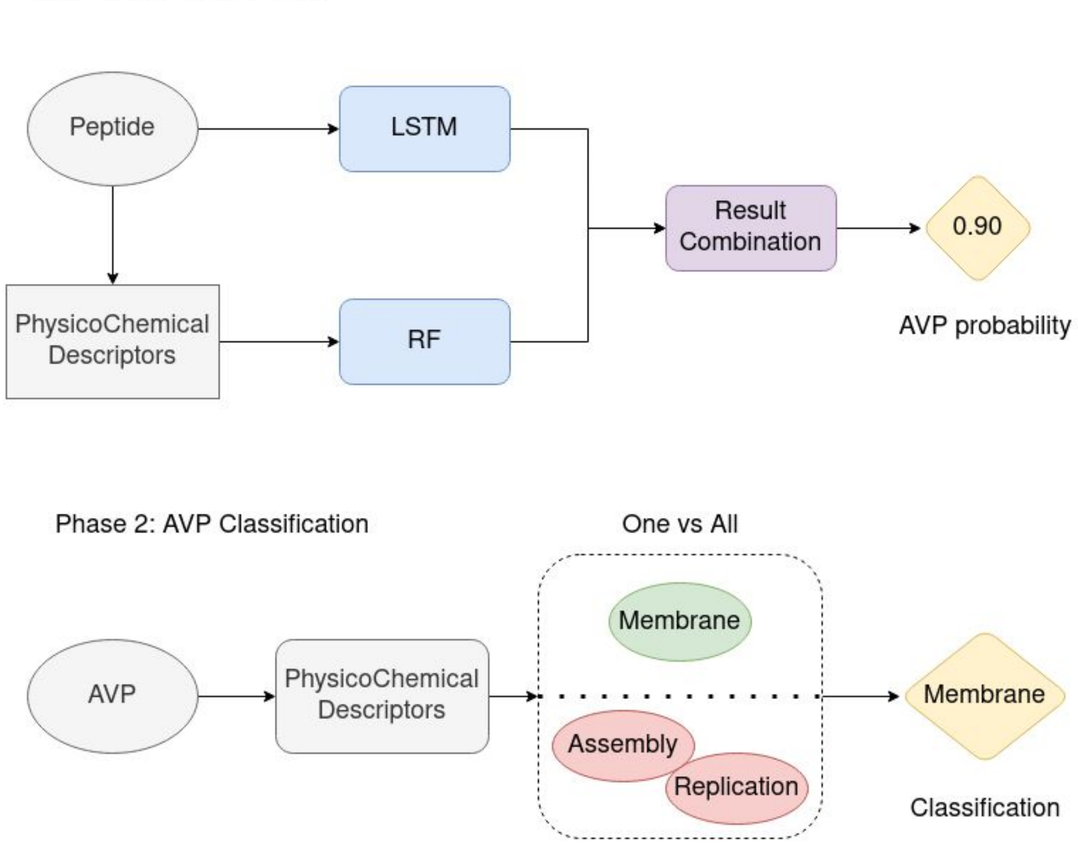

Fig. 1 - Arquitechture of the Two-Phase model in action. The first phase identifies compounds with antiviral activity and the second phase classifies those compounds based on their target.

RESULTS

We randomly the Dataset in three partitions: Training, Validation and Testing, with 70%, 15 and 15% of the sequences, respectively. The Training and Validation partitions were used for feature selection and hyperparameter setting (performed utilizing the *optuna* package). For this process we selected 4 possible feature sets for the RF module: Feature 1, Feature 2, Feature 3, Feature 4. A fifth “feature set” was only using the LSTM module (“Pure LSTM”). For each feature set 100 combinations of hyperparameters were tested.

We then trained the best performing model in the Training and Validation partitions and evaluated the performance on the Test partition. In the **AVP prediction task** the model had an **ACC of 95%** and an **AUC of 0.977.** Both results are shown bellow on figure 3.

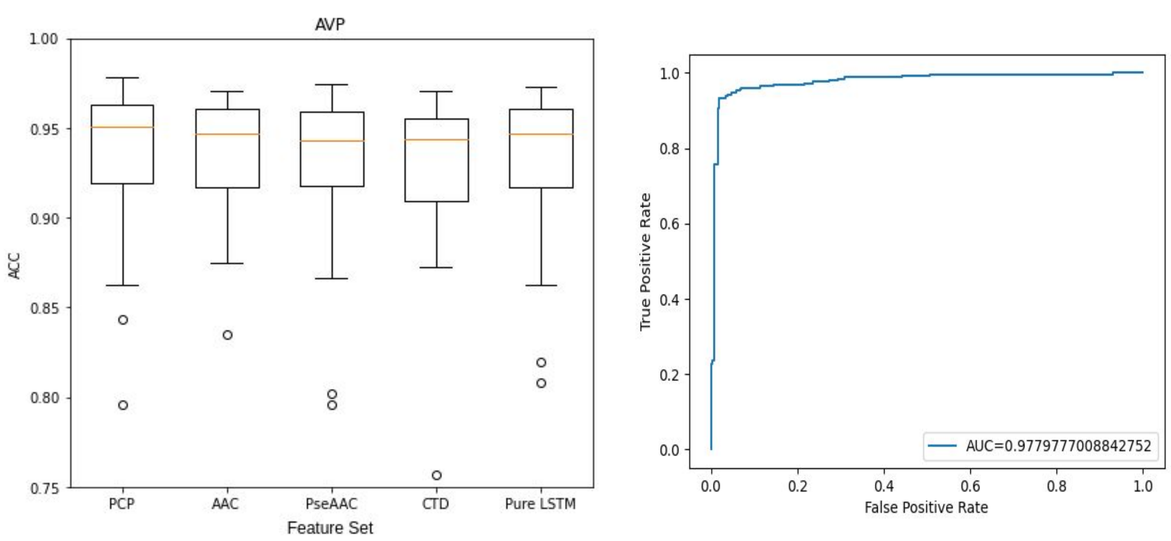

Fig. 3 - Boxplot with ACC for each trial in the hyperparameter and feature selection process and AVP prediction ROC curve.

Table 1 contains a comparison of our model with other available AVP prediction tools. In these tests we used the full set of AVP and non AMP peptide sequences gathered previously, before applying the homology reduction. We can see that our model outperformed the existing methods.

Table 1 - Comparison with other available tools.

Modelo	Treinamento	SubModelo	SENS	SPEC	ACC	CCM
DeepAVP	Thakur	-	0.967	0.9	0.933	0.87
AntiVPP	Thakur	-	0.87	0.97	0.93	0.87
Meta-iAVP	Thakur	-	0.917	**0.983**	0.949	0.9
AVPpred	Thakur	AVPphysico	0.933	0.917	0.925	0.85
Chang et al	Thakur	RFcompo	0.933	0.933	0.933	0.87
Chang et al	Thakur	RFcompo + agg	0.917	0.95	0.933	0.87
Chang et al	Thakur	RFphysico + structure + agg	0.95	0.867	0.908	0.82
EnAVPClass	Thakur	PseAAC+PCP	0.963	0.892	0.925	0.852
EnAVPClass	Local	PseAAC+PCP	**0.9711 ± 0.08**	0.9472 ± 0.22	**0.9584 ± 0.1**	**0.9175 ± 0.19**
AVPIden	Local	First-Stage	0.9243	0.9127	0.9185	-

Table 2 shows the performance of the classifier in identifying the correct activity class, based on the target of the AVP in the virus. We can see that the model shows grate results in both identifying AVPs, and in further helping characterize the kind of activity the peptides are expected to have.

Classe	Precisão	Recall	F1	ACC	suporte
Membrana	**0.91**	0.91	**0.91**	0.89	200
Replicação	0.87	0.86	0.86	0.89	138
Montagem	0.73	**0.92**	0.81	**0.98**	12

Table 2 - Performance in multiclass problem

CONCLUSIONS AND NEXT STEPS

- ✔ The proposed Model was able to outperform existing methods using the AVPs and Non-AVPs datasets, evidencing higher accuracy and specificity, but further tests and studies must be done.
- ✔ Our model was capable of identifying if a given amino acid sequence possesses the desired activity inside our dataset.
- ✔ Further experimental investigation and validation is required to validate if this technique can play an auxiliary role in drug discovery.
- ✔ The combination of Deep Learning and Machine Learning techniques, using different types of sequence characteristics, was shown to increase the performance of the model.

FINANCIAL SUPPORT

FAPESP
CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
CAPES