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.

PROCEEDINGS OF THE 31st European Symposium on Computer Aided Process Engineering (ESCAPE31), June 6-9, 2021, İstanbul, Turkey
M. Türkay, R. Gani (Editors).
http://dx.doi.org/10.1016/B978-0-323-88506-5.50312-0

Data-driven prediction of antiviral peptides based on periodicities of amino acid properties

Chris A. Kieslich*, Fatemeh Alimirzaei, Hyeju Song, Matthew Do, Paige Hall

[a]*Auburn University, Auburn, AL 36849 USA*

kieslich@auburn.edu

Abstract

With the emergence of new pathogens, e.g., methicillin-resistant Staphylococcus aureus (MRSA), and the recent novel coronavirus pandemic, there has been an ever-increasing need for novel antimicrobial therapeutics. In this work, we have developed support vector machine (SVM) models to predict antiviral peptide sequences. Oscillations in physicochemical properties in protein sequences have been shown to be predictive of protein structure and function, and in the presented we work we have taken advantage of these known periodicities to develop models that predict antiviral peptide sequences. In developing the presented models, we first generated property factors by applying principal component analysis (PCA) to the AAindex dataset of 544 amino acid properties. We next converted peptide sequences into physicochemical vectors using 18 property factors resulting from the PCA. Fourier transforms were applied to the property factor vectors to measure the amplitude of the physicochemical oscillations, which served as the features to train our SVM models. To train and test the developed models we have used a publicly available database of antiviral peptides (http://crdd.osdd.net/servers/avppred/), and we have used cross-validation to train and tune models based on multiple training and testing sets. To further understand the physicochemical properties of antiviral peptides we have also applied a previously developed feature selection algorithm. Future work will be aimed at computationally designing novel antiviral therapeutics based on the developed machine learning models.

Keywords: Computational biology, Machine learning, Support vector machines, Feature selection, Antiviral peptides

1. Introduction

With the increasing threat of viruses on human populations around the world, as evidenced by the recent COVID-19 pandemic, there is significant need for approaches for rapid development of treatments for novel viral outbreaks. If and when a new viral outbreak poses an eminent threat, the availability of tools for therapeutic design could enable the fast and efficient development of novel antiviral treatments. One promising class of antiviral treatments are anti-viral peptides (AVPs), which can act in a variety of ways, such as inhibiting replication, preventing binding to host cells, and interrupting virus-induced host signalling. Rational approaches have been previously used to successfully design AVPs, and more recent efforts have been aimed at using computational methods to predict their function based on the peptide sequence. One challenging aspect of developing machine learning models is identifying how to best encode or represent a peptide's sequence or properties. Most datasets of peptide function include peptides of varied length and to train a machine learning model one

must ensure that every peptide is represented by the same number of features. Past efforts for predicting AVPs have used peptide features that include the number of positively/negatively-charged amino acids, the charge of the peptide, the frequency of each amino acid in the sequence, the amount of possible H-bonds, molecular weight, and average hydrophobicity. One major limitation of these features is they do not maintain information regarding the ordering of amino acids or properties along the peptide structure, which is known to be crucial to protein structure and function. An alternative approach for developing physicochemical descriptors of protein sequences has been previously proposed that takes advantage of underlying periodicities in protein/peptide physicochemical properties (Eisenberg et al. 1984; Rackovsky 1998). Rackovsky (1998) has shown that periodicities of physicochemical properties along the sequence of a protein can be used to categorize families of protein structure/function. By using Fourier transforms and numerical tricks, it is possible to encode peptide sequences of varying lengths in terms of the same number of features based on the oscillation of amino acid properties.

In this work, we have used data analysis (i.e., PCA) and machine learning (i.e., support vector machines) to develop accurate models for predicting AVP sequences based on periodicities of amino acid properties. Additionally, by ranking the importance of the developed Fourier-based features, we were able to train SVM models with improved accuracy and generalizability, while also beginning to gain some insights into the importance of oscillations in physicochemical properties for AVP function.

2. Methods

In this work, we have used the R statistical language to perform all steps of our analysis including the generation of amino acid property factors using PCA, Fourier-based feature extraction, training/validating support vector machines, and feature selection. Below are more detailed descriptions of how these elements of our approach were implemented.

2.1. *AVP Dataset*

To develop data-driven classification models we need to have access to sufficiently large datasets, which contain both amino acid sequences and function labels. At present, there are multiple publicly available databases that hold the identities of some known antiviral peptides, including AVPpred (Thakur et al. 2012), APD3 (Wang et al. 2016), and CAMPR3 (Waghu et al. 2016). In the current study, we have focused on the AVPpred dataset, which contains 544 experimentally validated antiviral peptide sequences along with two sets of negative AVP: i) 407 experimentally validated nonactive peptides; and ii) 544 randomly selected non-secretory peptides. To eliminate the possibility of potential bias, we filtered the AVPpred dataset to eliminate any sequences with greater than 40% sequence identity. This was performed by first using the Clustal Omega webserver (https://www.ebi.ac.uk/Tools/msa/clustalo/) to align all sequences of the AVPpred dataset, and then applying hierarchical clustering in R to identify clusters of sequences sharing more than 40% sequence identity. The medoid of each cluster was selected as the representative sequence, resulting in 195 AVP sequences, 259 nonactive AVP sequences, and 492 randomly selected non-secretory peptides. The filtered dataset is what was used for all of the model training and validation in presented work.

2.2. *Principal component analysis*

To generate physicochemical Fourier-based features, we first need to convert amino acid sequences into numerical vectors based on amino acid properties. The AAindex dataset, found in the *protr* R package, is a collection of 544 amino acid properties from the literature that include various physicochemical descriptors. All of the 544 amino acid properties could be used to convert the amino acid sequences into property vectors; however, this would result in thousands of potential features once the property vectors were converted into Fourier coefficients. Alternatively, we can perform dimensionality reduction to reduce the number of amino acid property vectors prior to conversion to Fourier coefficients. In the current work, we have applied principal component analysis to generate amino acid property factors, as has been proposed previously. We used the *prcomp* function in R to extract principal components and to select a subset of the principal components based on contributions to the overall variance in the data (Figure 1).

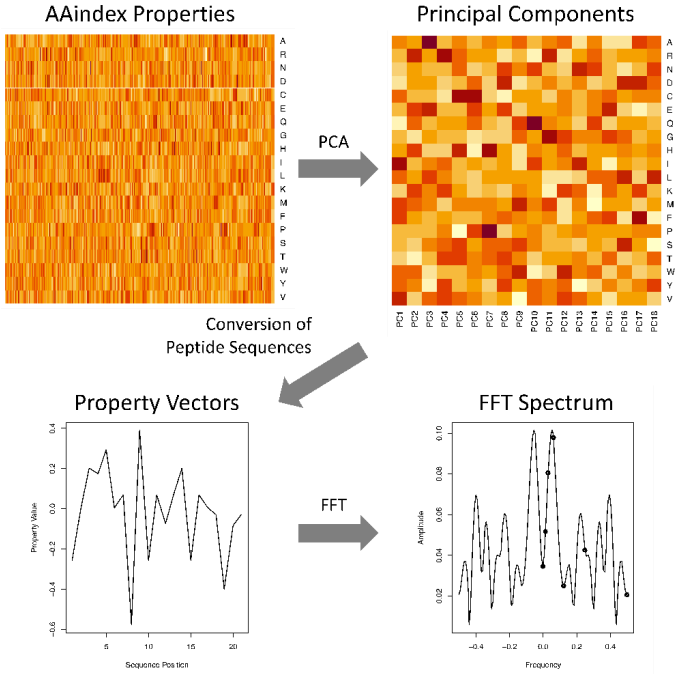

Figure 1. Schematic of feature extraction procedure based on amino acid property periodicities.

2.3. *Feature extraction*

Based on the generated amino acid property factors, we then converted the amino acid sequence of each peptide into physicochemical vectors (Figure 1). Fourier transforms, using the *fft* function in R, were then applied to each of the property factor vectors. To ensure that the same number of frequency (Fourier) components were generated for each property vector we used zero padding and an assumed maximum sequence length of 128 amino acids (maximum AVP sequence length in the dataset is 107 amino acids). The moduli of the complex Fourier coefficients for frequency values of 0, 0.015625, 0.03125, 0.0625, 0.125, 0.25, and 0.5 were selected as the features for training models (Figure 1). The frequency components (features) corresponding to periods that are

longer than a given peptide were set to zero. We eliminated features (columns) from the full set of features if more than 70% of peptides had a value of zero.

2.4. *Support vector machines*

All support vector machines were trained using the *svm* function of *e1071* R package based on the radial basis function nonlinear kernel. The cost and gamma hyperparameters of the SVM models were tuned using a grid-search with cost and gamma values based on powers of two, $2^n \ \forall\, n \in \{-9, \ldots, 8\}$, where n is an integer. Five-fold cross-validation, based on balanced training sets containing 435 AVP and 435 non-AVP sequences, was used to tune and validate the models based on first sorting the peptide sequences according to length and then select five training and testing sets with an equal number of samples for each peptide class. Model performance was measured based on classification accuracy and is reported as the fraction of classes (AVP or non-AVP) that was predicted correctly in the testing sets (Figure 3). The reported cross validation accuracies are the average of the classification accuracies for the five training and testing sets.

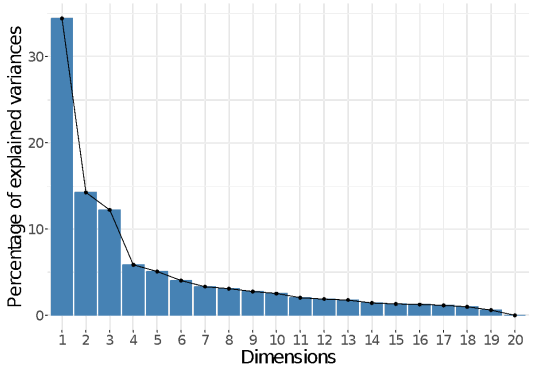

Figure 2. Variance explained by the 20 principal components from the PCA of the 544 AAindex amino acid properties.

2.5. *Feature selection*

Feature selection is a crucial aspect of data science as it can enable the identification of an essential set of predictive descriptors (features), as well as it can increase the robustness of models to prevent overfitting. Previously, we have developed a feature selection algorithm based on non-linear SVMs, which is general in nature and has been applied to predicting fault detection in chemical plants (Onel et al. 2018; Onel et al. 2019) and HIV-1 viral entry (Kieslich et al. 2016). The algorithm is model-based and requires first training a SVM model prior to computing a criterion that quantifies the contribution of each feature to the SVM objective function to determine which features to remove. The criterion (Eq. 1) is derived based on sensitivity analysis of the dual formulation of SVM models.

$$crit_k = -\frac{1}{2}\sum_i\sum_j \alpha_i^* \alpha_j^* y_i y_j \frac{\partial K\left(x_i \circ z, x_j \circ z\right)}{\partial z_k}\Big|_{z=1} \qquad (1)$$

The algorithm uses a greedy approach to rank the features, where we start with a training model based on all of the features, compute the criteria for all features, and remove a fraction of the features with the largest criteria values. In the presented work,

we removed 25% of the remaining features after each iteration of the algorithm and retuned the hyperparameters after each iteration of the algorithm. The feature ranking procedure was applied to each of the five training sets and a consensus ranking was generated based on the average rank of each feature across the five training sets.

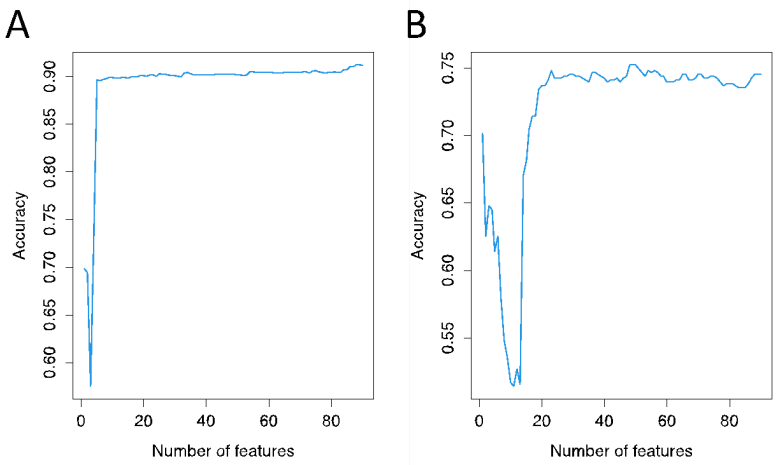

Figure 3. Feature contributions to model accuracy based on feature selection ranking. A) Feature selection results for classifying AVPs vs. random peptides; B) Feature selection results for classifying AVPs vs. non-AVPs.

3. Results

The principal component analysis of the amino acid properties from AAindex dataset was based on a data matrix consisting of 20 instances (amino acids) of 544 variables (properties). The PCA analysis generated 20 principal components that describe the more than 500 amino acid properties, and the contribution of each principal component is visualized in Figure 2. In choosing which principal components to use as amino acid property factors, we selected the principal components which contribute more than 1% of overall variance in the data. The first 18 principal components met our criteria (>1% variance) (Figure 2), and together describe over 99% of the total variance. The analysis resulted in 18 amino acid property factors that were used to generate 18 property vectors for each of the peptide sequences. For each peptide, the FFT spectrum of each property vector was computed and the frequency components corresponding to the sequence average and the oscillations with periods of 2, 4, 8, 16, 32, or 64 amino acids were extracted. This resulted in 126 features based on 7 frequency components from 18 property vectors for each sequence, which was filtered to 90 features by removing features with at least 70% of the values being zero.

Based on the generated features, we developed two SVM models, one to distinguish the AVP peptides from each of the types of non-AVP peptides (nonactive and random non-secretory peptides). For both classification tasks, we performed feature selection to rank the physicochemical features. To measure to the contribution of each feature to model accuracy we performed five-fold cross validation after adding each feature one at a time starting with the highest ranked feature. As can be seen in Figure 3, distinguishing AVPs from random non-secretory peptides is an easier classification task

than distinguishing AVPs from nonactive AVPs, since the maximum accuracy when using random peptides is 0.912 and only 0.752 for the nonactive peptides. Only 5 of the 90 features are necessary to achieve the majority of the accuracy of the AVP-vs-random model, while about 4 times as many features are necessary to achieve the maximum accuracy of the AVP-vs-nonactive model, which is further evidence of the difficulty distinguishing between active and nonactive AVPs.

4. Conclusions

In this study, we have developed support vector machine models that distinguish between antiviral peptide sequences and two classes of nonAVP sequences. To develop these models, we first generated amino acid property factors by applying principal component analysis to a dataset on amino acid properties from the literature. We then used the property factors to convert AVP sequences into property vectors that served as the input for Fourier analysis to extract the features used in training our models. The proposed approach for feature extraction and model development, including the incorporation of the feature selection algorithm, have potentially applications in prediction of peptide properties and function. Future work will be aimed at improving the Fourier-based encoding of peptide sequences and applying the approach to predicting various peptide functions/properties, as well as further development of approaches for SVM-based feature selection. The models developed in this study could have potential use in designing novel antiviral peptides but given the remaining challenges in distinguishing between active and nonactive AVPs further investigation is necessary, which may need to include both computational and experimental studies.

References

C.A. Kieslich, P. Tamamis, Y.A. Guzman, M. Onel, C.A. Floudas, 2016, Highly accurate structure-based prediction of HIV-1 coreceptor usage suggests intermolecular interactions driving tropism. PLOS ONE, 11(2), e0148974.

M. Onel, C.A. Kieslich, E.N. Pistikopoulos, 2019, A nonlinear support vector machine‐based feature selection approach for fault detection and diagnosis: Application to the Tennessee Eastman process. AIChE J., 65 (3), 992-1005.

M. Onel, C.A. Kieslich, Y.A. Guzman, C.A. Floudas, E.N. Pistikopoulos, 2018, Big Data Approach to Batch Process Monitoring: Simultaneous Fault Detection and Identification Using Nonlinear Support Vector Machine-based Feature Selection. Comput. Chem. Eng., 115, 46-63.

S. Rackovsky, 1998, Hidden sequence periodicities and protein architecture, P. Natl. Acad. Sci. USA, 95, 8580-8584.

N. Thakur, A. Qureshi, M. Kumar, 2012, AVPpred: collection and prediction of highly effective antiviral peptides, Nucleic Acids Res., 40, W199- W204.

D. Eisenberg, R.M. Weiss, T.C. Terwilliger, 1984, The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity, P. Natl. Acad. Sci. USA, 81(1), 140-44.

F.H. Waghu, R.S. Barai, P. Gurung, S. Idicula-Thomas, 2016, CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nuc. Acids Res., 4, 44,1094-1097.

G. Wang, X. Li, Z. Wang, 2016, APD3: the antimicrobial peptide database as a tool for research and education, Nuc. Acids Res., 4, 44,1087-93.