Characterization and Prediction of Dengue Virus targeting peptides based on three class of descriptors using k-NN and Random Forest algorithm.

Elakkiya Elumalai
Pondicherry University https://orcid.org/0000-0001-9449-6462
Suresh Kumar Muthuvel (muthuvels@hotmail.com)
Pondicherry University https://orcid.org/0000-0001-7083-6565

Research Article

Keywords: Dengue Virus, Enhanced amino acid composition, physiochemical property, random forest, machine learning, cross validation

Posted Date: May 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1652354/v1

Abstract

Dengue virus peptides are emerging as potential therapeutics for dengue infection. Due to the important role of dengue peptides in curbing dengue infection, their identification has proven crucial in terms of infection biology. To calculate differences between amino acids and physiochemical attributes, statistical tests and F-scores were used in this work. In this study, we presented the first evidence, to our knowledge, for the relationship between dengue virus inhibiting and non-inhibiting peptides with amino acid use and biological properties. We found that the frequency of Glycine (G), Phenylalanine (F), and Tryptophan (W) was significantly higher in dengue virus inhibitory peptides. Similarly, aromatic amino acids in non-inhibiting peptides were found to be less than 5%. The distribution of solvent accessible residues in non-inhibiting peptides was found to be less as compared to inhibiting peptides. The alpha helices and beta sheets in dengue virus inhibiting peptides are equally distributed but in non-inhibiting peptides, the proportion of beta sheets is more as compared to alpha helices. We have used 8 machine learning algorithms to predict dengue peptides. Here we have used three class of descriptors namely Amino acid composition (AAC), grouped amino acid composition, transition and distribution (GAAC) and Composition, transition and distribution features (CTDC). We have compared all 8 machine learning models for the three classes. The machine learning algorithm were, Random forest (RF), Multi-layer perceptron (MLP), Support Vector Machine (SVM), Logistic regression, K-Nearest Neighbour (k-NN), Naivebayes, Adaboost and Bagging. The best model was reposted as AAC_k-NN_model, accuracy was 90.47. Subsequently, other four best model were AAC_SVM_model, CTDC_SVM_model, AAC_RF_model and GAAC_RF_model. In all the three class, k-NN and RF models were found to be the best classifier. Our classifier produced superior predicting outcomes when compared to previously developed algorithms. In conclusion, we looked at the differences in amino acids and physiochemical properties between dengue viral peptides, using the grouped amino acid composition to build a classifier that predicts these dengue virus inhibitory peptide.

Highlights

- Amino acid content, grouped amino acid content, CTDC descriptors were used on 8 machine learning algorithms to predict dengue virus inhibiting peptides.
- The k-NN, SVM and RF were found to be the best model for classifying dengue inhibiting and non-inhibiting peptides.
- Frequency of Glycine (G), Phenylalanine (F), and Tryptophan (W) was significantly higher in dengue virus inhibitory peptides.
- Aromatic amino acids in non-inhibiting peptides were found to be less than 5%.
- In non-inhibiting peptides, the distribution of solvent accessible residues was found to be less than in inhibiting peptides.
- The testing accuracy of the best models AAC_RF and AAC_k-NN_model was 85.71 %.

1. Introduction

Dengue virus (DENV) is the mosquito-borne flavivirus that frequently infects people in subtropical and tropic areas. As per the reports of the World Health Organization, over 40% of the world's population are at risk of dengue infection [1]. Dengue virus infections cause severe illness, known as dengue haemorrhagic fever (DHF). It is majorly characterized by vascular leakage, which further develops into life-threatening dengue shock syndrome (DSS) [2]. It leads to high mortality of DHF/DSS. DENV NS1 is a 48-kDa glycoprotein that is highly conserved among all flaviviruses [3]. NS1 is essential for viral replication and immune evasion [4][5]. The triggering hyperpermeability of human endothelial cells in-vitro and systemic vascular leakage in-vivo is caused by the pathogenic effect of secreted DENV non-structural protein 1 (NS1) [6]. The NS1 disrupts endothelial glycocalyx layer (EGL), inducing the shedding of heparan sulfate glycoprotein and degradation of sialic acid. It has been shown that NS1 activates cathepsin L which activates heparanase via enzymatic cleavage. This enzyme act on the breakdown of heparan sulfate proteoglycans. Therefore, DENV patients have high heparan sulfate and sialic acid in their serum [7].

The use of peptides as therapeutic agents for DENV infection has previously been investigated. As competitive inhibitors of virus entrance and replication, these peptides were engineered to disrupt active regions of viral proteins or to imitate specific sections of viral proteins. Peptide inhibitors have been shown to target viral structural proteins C, prM, and E, as well as viral NS1, NS2B/NS3 protease, and NS5 methyltransferase during DENV infection. [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

Dengue virus peptides are emerging as potential therapeutics for dengue infection. Due to the important role of dengue peptides in curbing dengue infection, their identification has proven crucial in terms of infection biology. To calculate differences between amino acids and physiochemical attributes, statistical tests and F-scores were used in this work.

Here, we have proposed best classification algorithm to classifiy dengue virus inhibiting peptides using three main descriptors namely; Amino Acid content, grouped amino acid content and CTDC. The binary dataset for developing machine learning model were taken from literatures and dengue peptides-oriented databases. We have used 8 machine learning algorithms to predict dengue peptides. Here we have used three class of descriptors namely Amino acid composition (AAC), grouped amino acid composition, transition and distribution (GAAC) and Composition, transition and distribution features (CTDC). We have compared all 8 machine learning models for the three classes. The machine learning algorithm were, Random forest (RF), Multi-layer perceptron (MLP), Support Vector Machine (SVM), Logistic regression, K-Nearest Neighbour (k-NN), Naivebayes, Adaboost and Bagging. The best model was reposted as AAC_k-NN_model, accuracy was 90.47. Subsequently, other four best model were AAC_SVM_model, CTDC_SVM_model, AAC_RF_model and GAAC_RF_model. In all the three class, k-NN and RF models were found to be the best classifier. Our classifier produced superior predicting outcomes when compared to previously developed algorithms. In conclusion, we looked at the differences in amino acids and physiochemical properties between dengue viral peptides, using the grouped amino acid composition to build a classifier that predicts these dengue virus inhibitory peptide.

2. Materials And Methods

2.1. Dataset

In this study, Dengue virus inhibiting peptides were downloaded from the AVPdb, a database of antiviral peptides that have been experimentally confirmed against medically significant viruses [20], which consisted of 89 dengue virus inhibiting peptides. The 11 peptides were taken from a paper entitled "**Peptides targeting dengue viral nonstructural protein 1 inhibit dengue virus production**". The negative dataset was taken from AVPdb Database [19]. All the peptide sequences were checked in Cluster Database at High Identity with Tolerance (CD-HIT) [21] in order to generate a high-quality dataset for this research. Finally, we have categorized our both dataset into training and testing with 7:3 ratio.

2.2 Descriptor selection

We selected three descriptors. 1- Amino acid content (AAC) which calculates amino acid frequency in peptide sequence.2- Grouped Amino Acid Composition (GAAC), twenty amino acids are categorized into five classes (aliphatic, aromatic, positive, negative, uncharge). It calculates the frequency of each class. 3- The composition, transition and distribution (CTDC) features represent amino acid distribution patterns of a specific structural or physiochemical property in a peptide sequence. We used iLearnplus Web [22] for descriptor selection and machine learning model development.

2.3 Clustering and dimensionality reduction

The three descriptor's data were used as input for clustering. K-means clustering was used with cluster size of 2. The basic idea is to initialize cluster centers, move each point to its new nearest center and calculating the mean of the member points to update the clustering centers and repeat the process until the convergence [23].

The Principal Component Analysis (PCA) is used to describe useful variants [24]. The data was used for principal component analysis for dimensionality reduction. The main three principal components were retrieved. The dimensionality reduction data was used as input for feature selection and normalization.

2.4 Feature selection and normalization

F score is used for class discrimination. F-score can measure the discrimination between sets of real numbers [25]. For feature selection, F score value was used and 10 best features was found. The values of features were transformed into three principal components. The features were normalized using Z Score. Nowadays, microarrays data also being normalized using Z score [26].

2.5 Machine learning algorithms used

A big part of machine learning is classification — we want to know what class a new peptide is (Dengue inhibiting peptide or non-inhibiting). We have considered 8 machine learning algorithms with the following parameters.

Random forest parameters:

Tree number:100; number of thread: 1; Autooptimization Tree range from: 50; Tree range to: 500; step:50; cross-validation:5

Random forest is a Supervised Machine Learning Algorithm which builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It can handle continuous variable as well as categorical variables. It has shown good results in classification problems [27].

Light Gradient Boosting Machine (LightGBM) parameters:

Boosting type: gbdt; number of leaves:31; maximum depth: -1; learning rate: 0.1; number of threads:1 Auto optimization, leaves range: 20:100:2;depth range: 15:55:10; learning rate range: 0.01,0.15,0.02

It is a gradient boosting framework that makes use of tree-based learning algorithms which performs quite well in huge dataset. It is an extremely fast and accurate classifier, employed for binary classification of Biological sequences [28].

Support Vector Machine (SVM):

Kernel: rbf; Penalty:1, Gamma: Auto, Penalty from:1'Penalty to: 15'Gamma from: -10,Gamma to: 5

SVM is one of the most popular Supervised Learning algorithms, which is primarily used for Classification. It creates multiple decision boundary that can separate data points in n-dimensional space into classes. The decision boundary is determined by extreme vectors called a support vector. The best decision boundary is called hyperplane. Hence, this algorithm is called as Support Vector Machine. The SVM is widely used in classification of biological sequences [29].

Logistic regression (LR):

None

Logistic regression is a supervised learning classification algorithm which primarily classify two classes. It is used to predict the probability of a target variable. Therefore, the LR is also used in classification of biological sequences [30].

k-Nearest Neighbour (k-NN):

Top k values:3

The k-NN is also Supervised Learning technique which calculates the similarity between the query data with available dataset. It classifies the query data based on similarity percentage. The k-NN algorithm can also be used for regression. The k-NN is widely used for biological image and sequence classification [31].

Naïve Bayes

None

Naive Bayes classifiers are based on Bayes' Theorem. It is a collection of classification algorithm. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. The Naïve Bayes is also used in classification of biological sequences which helps in taxonomy [32].

Adaptive Boosting (Adaboost)

None

This method is based on the principle that learners growing sequentially. The weak learners are converted into strong learner. It is boosting technique used in which the weights are re-assigned to each instance i.e., higher weights assigned to incorrectly classified instances. It is an ensemble method in Machine Learning. This algorithm is generally used in biological sequence classification [33].

MLP (Multi-Layer Perceptron)

Hidden layer Size: 32,32

Epochs: 200

Activation: relu

Optimizer: adam

Multi layer perceptron (MLP) is a supplement of feed forward neural network which consists of three layers namely a input layer, arbitrary number of hidden layers and a output layer. The input signal is handled by input layer. The hidden layers are the true computational engine of MLP. The output layer helps in prediction and classification. The data is processed in the forward direction from input to output layer. The nodes in the MLP are trained with the back propagation learning algorithm. MLP can solve non-linear problems. MLP is used in biological sequence classification [34].

Bagging:

N_estimators: 10

Number of CPU: 1

Bagging is an ensemble learning technique that enhances the accuracy and performance of machine learning algorithms. It avoids overfitting of data and is used for both regression and classification models, specifically for decision tree algorithms. Bagging algorithm is rarely used in biological sequence classification [35].

Here, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions [36]. The normalized dataset (Training set: 102,3; Testing set: 14,3) was taken as input and loaded in these machine learning algorithms. The cross validation was set to 5 in all cases. Subsequently, 24 models were developed and compared with respect to model Accuracy, ROC and PRC. The ROC and PRC curve was ploted. The evaluation metrics was reported.

2.6 Best 5 model validation in testing sequences

The best five models were validated with testing sequences. The prediction results were tabulated. We also tested best model in five random peptde sequences.

3. Results And Discussion

3.1 Dataset

As per the protocol of iLearnWeb Plus server, we annotated all sequences for classification. The protocol for writing sequence is given below.

>name|class|category

Sequence

***All the sequences appended at the end**

Here, we can give any name (alphanumeric with underscore). we had two class; 1 for dengue virus inhibiting peptides, 0 for dengue virus non-inhibiting peptides. Totally, we collected 100 experimentally validated dengue virus inhibiting peptides. Here, category means training and testing dataset. We split the sequences into training and testing set in 7:3 ratio. Similarly, we had 16 negative datasets. This set also we split into 7:3 ratio. **All the sequences appended at the end.**

3.2 Descriptor generation and data distribution

We generated descriptors for all 116 peptides. The generated 20 descriptors under AAC are given in supplementary table 1. This numeric value indicates frequency of Amino acid in peptides. In AAC, Tryptophan frequency differentiates dengue virus inhibiting peptides from non-inhibiting peptides. In non-inhibiting peptides the occurrence of tryptophan is almost 0. In various literatures, it has been shown that tryptophan is very important for delivering antimicrobial activity [37, 38]. Similarly, Glycine, tryptophan and phenylalanine frequency in non-inhibiting peptide is comparatively less than inhibiting peptides. (Table 1) and it is well supported by published article [39]. The generated 5 descriptors under GAAC are given in supplementary table 2. In GAAC, Aromatic amino acids in non-inhibiting peptides were found to be less than 5%. It has been reported that aromatic amino acids play a vital role in viral defense [40]. The generated 39 descriptors under CTDC are given in supplementary table 3. In CTDC, the distribution of solvent accessible residues in non-inhibiting peptides was found to be less. The alpha helices and beta

sheets in dengue virus inhibiting peptides are equally distributed but in non-inhibiting peptides, the proportion of beta sheets is more as compared to alpha helices.

The alpha helical content in peptides determine its antiviral activity [41]. The data distribution for AAC, GAAC and CTDC is given in Fig. 1.

3.3 Machine learning model and its validation

The amino acid composition of a protein has been widely utilized for the prediction of peptide categories [42–52]. All descriptors under AAC, GAAC and CTDC was used for clustering (Fig. 2) and dimensionality reduction (Fig. 3).

5.3.3 Machine learning model and its validation

The top 10 features were selected and transformed into 3 three principal components. Further, principal component values for each sequence were normalized. The normalized data for AAC, GAAC and CTDC is shown in Fig. 4.

The normalized data was used as input for model development. The RF algorithm is widely used for better understanding and prediction of antiviral peptides [53]. We have compared all 8 machine learning models for the three class. The models were saved in pickle (pkl) format. The machine learning algorithm were, Random forest (RF), Multi-layer perceptron (MLP), Support Vector Machine (SVM), Logistic regression, K-Nearest Neighbour (k-NN), Naivebayes, Adaboost and Bagging. The best model was reposted as AAC_k-NN_model, accuracy was 90.47. Subsequently, other four best model were AAC_SVM_model, CTDC_SVM_model, AAC_RF_model and GAAC_RF_model (Table 1).

Table 1
Best five model's metrics on training data

Id	Sn	Sp	Pre	Acc	MCC	F1	AUROC	AUPRC
AAC_k-NN_model	96.844	61	92.362	90.47	0.634	0.9443	0.8992	0.9777
AAC_SVM_model	96.842	56	91.618	89.6	0.564	0.9393	0.8821	0.9598
CTDC_SVM_model	95.788	46	89.666	87.066	0.4763	0.9239	0.9084	0.9718
AAC_RF_model	95.788	43	88.512	86.268	0.4465	0.9191	0.8642	0.9653
GAAC_RF_model	92.632	56	91.212	86.268	0.5091	0.9163	0.8671	0.9675

The ROC and PRC curve for all the best five models are shown in Fig. 5. In this table, best five models have been reported.

The boxplot for all models with 8 different evaluation parameters is shown in Fig. 6.

On looking into the model metric, we selected all the best 5 models for its testing. These models were tested in testing sequences and new sequences. The testing metrics have been tabulated below in Table 2.

Table 2
Best five model's metrics on testing data

Id	Sn	Sp	Pre	Acc	MCC	F1	AUROC	AUPRC
AAC_k-NN_model	100	50	83.33	85.71	0.6455	0.9091	0.975	0.9917
AAC_SVM_model	100	0	71.43	71.43	0	0.833	0.175	0.6184
CTDC_SVM_model	100	0	71.43	71.43	0	0.833	0.475	0.6529
AAC_RF_model	100	50	83.33	85.71	0.6455	0.9091	0.975	0.9905
GAAC_RF_model	100	25	76.92	78.57	0.4385	0.8696	0.6	0.8252

The testing metrics indicates the reliability of two model in peptide classification. The AAC_k-NN_model and AAC_RF_model have an accuracy of 85.71%. These models were checked for classifying a random five peptide sequences.

> 1

MDPPPPKKKK

> 2

ECCCCFFFK

> 3

PRDCEAK

> 4

IAGIDH

> 5

RHKFDPR

Table 3
AAC_RF_model classifying five sequences.

Score_0	Score_1
0.408	0.592
0.02	0.98
0.484	0.516
0.256	0.744
0.2	0.8

Table 4
AAC_k-NN_model classifying five sequences.

Score_0	Score_1
0.666667	0.333333
0	1
0.333333	0.666667
0	1
0.333333	0.666667

Here, score 0 is score for non-inhibiting peptides. Score 1 is score for inhibiting peptides. Here, we can see clearly except sequence 1, both models are predicting remaining sequences as dengue inhibiting peptides. There is a conflict for sequence 1 in both models.

The AAC descriptors with RF and k_NN perform wells in new sequences. The AAC_RF_model is well correlated with CTDC_RF_model. The correlation coefficient is 0.9996. Similarly, the AAC_k-NN_model is well correlated with CTDC_k_-NN_model. The correlation coefficient is 0.9910 (Fig. 7).

The successful predictive performance obtained in our study clearly demonstrated that the AAC descriptors with Random Forest and k-NN were quite suitable for predicting these peptides inhibiting dengue virus The model was evaluated on testing data and five new random sequences. Therefore, Compared to the GAAC and CTDC which decreases information redundancy, overfitting and simplifies the protein complexity, AAC is good. To determine which amino acids and biological features were most discriminative between dengue virus inhibiting and non-inhibiting peptides, we analyzed differences in amino acids and biological properties. We aimed to create a classifier that could predict dengue virus inhibitory peptides based on AAC. As a result, these descriptors served as RF's and k-NN input parameters.

4. Conclusion

There is currently no effective dengue virus (DENV) therapeutic. In this study, we presented the first evidence, to our knowledge, for the relationship between dengue virus inhibiting and non-inhibiting peptides with amino acid use and biological properties. We found that the frequency of Glycine (G), Phenylalanine (F), and Tryptophan (W) was significantly higher in dengue virus inhibitory peptides. Similarly, aromatic amino acids in non-inhibiting peptides were found to be less than 5%. The distribution of solvent accessible residues in non-inhibiting peptides was found to be less as compared to inhibiting peptides. The alpha helices and beta sheets in dengue virus inhibiting peptides are equally distributed but in non-inhibiting peptides, the proportion of beta sheets is more as compared to alpha helices. We applied 8 machine learning algorithms with three class of descriptors to predict dengue virus inhibiting peptides. The successful predictive performance obtained in our study clearly demonstrated that AAC descriptors combined with RF and k-NN were quite suitable for predicting these two peptide categories. The AAC_RF and AAC_k-NN model has improved accuracy of 85.71%. Based on these data, we believed that our classifier, may facilitate dengue virus inhibition peptide prediction.

Declarations

Acknowledgements

We sincerely acknowledge the Centre for Bioinformatics, for providing computational facility to carry out this research work.

Funding

This project was supported by the Indian Council of Medical Research (ICMR, New Delhi).

The grant number is No: **45/36/2019-PHA/BMS**

Ethics declarations: Not applicable as we didn't do any experiments using model organism.

Competing interests

The authors declare that they have no competing interests.

Availability of data and material: Yes, we uploaded all data and material while submission

Code availability: Not Applicable

Author information

Suresh Kumar Muthuvel designed this work and Elakkiya Elumalai performed analysis and wrote the manuscript.

Affiliations

Center for Bioinformatics, Pondicherry University, Pondicherry, India

Elakkiya Elumalai & Suresh Kumar Muthuvel

Contributions

SKM designed the experiments. EE performed the experiments. EE analyzed the data. EE wrote the manuscript. SKM proofed the manuscript. Both authors read and approved the final manuscript.

Corresponding author

Correspondence to Suresh Kumar Muthuvel.

Consent to participate: This is a computational biology work so consent is not required.

Consent for publication: I, **Prof. Suresh Kumar Muthuvel**, undersigned, give my consent for the publication of identifiable details, which can include photograph(s) and/or videos and/or case history and/or details within the text (“Material”) to be published in the Journal and Article. Therefore, anyone can read material published in the Journal.

References

1. Bhatt, S.; Gething, P.W.; Brady, O.J.; Messina, J.P.; Farlow, A.W.; Moyes, C.L.; et al. The global distribution and burden of dengue. Nature. **2013**;496(7446):504–7.
2. World Health Organization. Dengue haemorrhagic fever: diagnosis, treatment, prevention and control, 2nd ed. World Health Organization **1997**.
3. Muller, D.A.; Young, P.R.; The flavivirus NS1 protein: molecular and structural biology, immunology, role in pathogenesis and application as a diagnostic biomarker. Antiviral Res. **2013**;98(2):192–208.
4. Uno, N.; Ross, T.M.; Dengue virus and the host innate immune response. Emerg Microbes Infect. **2018**;7(1):167.
5. Youn, S.; Li, T.; McCune, B.T.; Edeling, M.A.; Fremont, D.H.; Cristea, I.M., et al. Evidence for a genetic and physical interaction between nonstructural proteins NS1 and NS4B that modulates replication of West Nile virus. J Virol. **2012**;86(13):7360–71.
6. Puerta-Guardo, H.; Glasner, D.R.; Harris, E.; Dengue virus NS1 disrupts the endothelial glycocalyx, leading to hyperpermeability. PLoS Pathog. **2016**;12(7):e1005738.
7. Tang, TH-C; Alonso, S.; Ng, LF-P; Thein, T-L; Pang, VJ-X; Leo, Y-S; et al. Increased serum hyaluronic acid and heparan sulfate in dengue fever: Association with plasma leakage and disease severity. Sci Rep. **2017**;7:46191.
8. Tambunan, U. S. & Alamudi, S. Designing cyclic peptide inhibitor of dengue virus NS3-NS2B protease by using molecular docking approach. Bioinformation 5, 250–254 (2010).

9. Lok, S.-M. et al. Release of dengue virus genome induced by a peptide inhibitor. PLoS ONE 7, e50995 (2012).
10. Tambunan, U. S., Zahroh, H., Utomo, B. B. & Parikesit, A. A. Screening of commercial cyclic peptide as inhibitor NS5 methyltransferase of dengue virus through molecular docking and molecular dynamics simulation. Bioinformation 10, 23–27 (2014).
11. Li, L. et al. Structure-guided Discovery of a Novel Non-peptide Inhibitor of Dengue Virus NS2B-NS3 Protease. Chem. Biol. Drug Des. 86, 255–264 (2015).
12. Panya, A. et al. A peptide inhibitor derived from the conserved ectodomain region of DENV membrane (M) protein with activity against dengue virus infection. Chem. Biol. Drug Des. 86, 1093–1104 (2015).
13. da Silva-Junior, E. F. & de Araujo-Junior, J. X. Peptide derivatives as inhibitors of NS2B-NS3 protease from Dengue, West Nile, and Zika flaviviruses. Bioorg. Med. Chem. 27, 3963–3978 (2019).
14. Faustino, A. F. et al. Structural and functional properties of the capsid protein of dengue and related flavivirus. Int. J. Mol. Sci. 20, e3870 (2019).
15. Ji, M. et al. An antiviral peptide from *Alopecosa nagpag* spider targets NS2B-NS3 protease of flaviviruses. Toxins (Basel) 11, 584 (2019)
16. Zhu, T. et al. Development of peptide-based chemiluminescence enzyme immunoassay (CLEIA) for diagnosis of dengue virus infection in human. Anal. Biochem. 556, 112–118 (2018).
17. Isa, D. M. et al. Dynamics and binding interactions of peptide inhibitors of dengue virus entry. J. Biol. Phys. 45, 63–76 (2019).
18. Behnam, M. A. M., Nitsche, C., Vechi, S. M. & Klein, C. D. C-Terminal residue optimization and fragment merging: discovery of a potent peptide-hybrid inhibitor of dengue protease. ACS Med. Chem. Lett. 5, 1037–1042 (2014).
19. Songprakhon P, Thaingtamtanha T, Limjindaporn T, et al. Peptides targeting dengue viral nonstructural protein 1 inhibit dengue virus production. Sci Rep. 2020;10(1):12933. Published 2020 Jul 31. doi:10.1038/s41598-020-69515-9
20. Qureshi A, Thakur N, Tandon H, Kumar M. AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. Nucleic Acids Res. 2014 Jan;42(Database issue):D1147-53. doi: 10.1093/nar/gkt1191. Epub 2013 Nov 26. PMID: 24285301; PMCID: PMC3964995.
21. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010 Mar 1;26(5):680-2. doi: 10.1093/bioinformatics/btq003. Epub 2010 Jan 6. PMID: 20053844; PMCID: PMC2828112.
22. https://ilearnplus.erc.monash.edu/
23. Zhang Y, Liu N, Wang S. A differential privacy protecting K-means clustering algorithm based on contour coefficients. PLoS One. 2018 Nov 21;13(11):e0206832. doi: 10.1371/journal.pone.0206832. PMID: 30462662; PMCID: PMC6248925.

24. David CC, Jacobs DJ. Principal component analysis: a method for determining the essential dynamics of proteins. Methods Mol Biol. 2014;1084:193-226. doi: 10.1007/978-1-62703-658-0_11. PMID: 24061923; PMCID: PMC4676806.
25. Sokolova M., Japkowicz N., Szpakowicz S. (2006) Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: Sattar A., Kang B. (eds) AI 2006: Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science, vol 4304. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11941439_114
26. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. J Mol Diagn. 2003 May;5(2):73-81. doi: 10.1016/S1525-1578(10)60455-2. PMID: 12707371; PMCID: PMC1907322.
27. Denisko D, Hoffman MM. Classification and interaction in random forests. Proc Natl Acad Sci U S A. 2018 Feb 20;115(8):1690-1692. doi: 10.1073/pnas.1800256115.
28. Sharma A, Singh B. AE-LGBM: Sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and LightGBM. Comput Biol Med. 2020 Oct;125:103964
29. Meng C, Jin S, Wang L, Guo F, Zou Q. AOPs-SVM: A Sequence-Based Classifier of Antioxidant Proteins Using a Support Vector Machine. Front Bioeng Biotechnol. 2019 Sep 18;7:224. doi: 10.3389/fbioe.2019.00224.
30. Nepal R, Spencer J, Bhogal G, Nedunuri A, Poelman T, Kamath T, Chung E, Kantardjieff K, Gottlieb A, Lustig B. Logistic regression models to predict solvent accessible residues using sequence- and homology-based qualitative and quantitative descriptors applied to a domain-complete X-ray structure learning set. J Appl Crystallogr. 2015 Nov 10;48(Pt 6):1976-1984. doi: 10.1107/S1600576715018531.
31. Dongardive, J., Abraham, S. (2016). Protein Sequence Classification Based on N-Gram and K-Nearest Neighbor Algorithm. In: Behera, H., Mohapatra, D. (eds) Computational Intelligence in Data Mining—Volume 2. Advances in Intelligent Systems and Computing, vol 411. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2731-1_15
32. Ziemski M, Wisanwanichthan T, Bokulich NA, Kaehler BD. Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences. Front Microbiol. 2021 Jun 18;12:644487.
33. Wang Y, Ru J, Jiang Y, Zhang J. Adaboost-SVM-based probability algorithm for the prediction of all mature miRNA sites based on structured-sequence features. Sci Rep. 2019 Feb 6;9(1):1521. doi: 10.1038/s41598-018-38048-7. PMID: 30728425; PMCID: PMC6365589.
34. Li Y, Zhang Z, Teng Z, Liu X. PredAmyl-MLP: Prediction of Amyloid Proteins Using Multilayer Perceptron. Comput Math Methods Med. 2020 Nov 20;2020:8845133. doi: 10.1155/2020/8845133.
35. Govindan G, Nair AS. Bagging with CTD--a novel signature for the hierarchical prediction of secreted protein trafficking in eukaryotes. Genomics Proteomics Bioinformatics. 2013 Dec;11(6):385-90. doi: 10.1016/j.gpb.2013.07.005.
36. Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front Aging Neurosci. 2017;9:329. Published 2017

Oct 6. doi:10.3389/fnagi.2017.00329

37. Chan DI, Prenner EJ, Vogel HJ. Tryptophan- and arginine-rich antimicrobial peptides: structures and mechanisms of action. Biochim Biophys Acta. 2006 Sep;1758(9):1184-202. doi: 10.1016/j.bbamem.2006.04.006. Epub 2006 Apr 21. PMID: 16756942.
38. Pasupuleti M, Chalupka A, Mörgelin M, Schmidtchen A, Malmsten M. Tryptophan end-tagging of antimicrobial peptides for increased potency against Pseudomonas aeruginosa. Biochim Biophys Acta. 2009 Aug;1790(8):800-8. doi: 10.1016/j.bbagen.2009.03.029. Epub 2009 Apr 5. PMID: 19345721.
39. Sala A, Ardizzoni A, Ciociola T, Magliani W, Conti S, Blasi E, Cermelli C. Antiviral Activity of Synthetic Peptides Derived from Physiological Proteins. Intervirology. 2018;61(4):166-173. doi: 10.1159/000494354. Epub 2019 Jan 17. PMID: 30654366.
40. Sitaram N. Antimicrobial peptides with unusual amino acid compositions and unusual structures. Curr Med Chem. 2006;13(6):679-96. doi: 10.2174/092986706776055689. PMID: 16529559.
41. Cho NJ, Dvory-Sobol H, Xiong A, Cho SJ, Frank CW, Glenn JS. Mechanism of an amphipathic alpha-helical peptide's antiviral activity involves size-dependent virus particle lysis. ACS Chem Biol. 2009 Dec 18;4(12):1061-7. doi: 10.1021/cb900149b. PMID: 19928982.
42. Chen YL, Li QZ. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J Theor Biol 2007; 248(2): 377-81.http://dx.doi.org/10.1016/j.jtbi.2007.05.019 PMID: 17572445
43. Chen YL, Li QZ. Prediction of the subcellular location of apoptosis proteins. J Theor Biol 2007; 245(4): 775-83.http://dx.doi.org/10.1016/j.jtbi.2006.11.010 PMID: 17189644
44. Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 2002; 277(48): 45765-9. http://dx.doi.org/10.1074/jbc.M204161200 PMID: 12186861
45. Chou KC, Elrod DW. Bioinformatical analysis of G-proteincoupled receptors. J Proteome Res 2002; 1(5): 429-33. http://dx.doi.org/10.1021/pr025527k PMID: 12645914
46. Cai YD, Ricardo PW, Jen CH, Chou KC. Application of SVM to predict membrane protein types. J Theor Biol 2004; 226(4): 373-6. http://dx.doi.org/10.1016/j.jtbi.2003.08.015 PMID: 14759643
47. Mondal S, Bhavna R, Mohan Babu R, Ramakumar S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 2006; 243(2): 252-60. http://dx.doi.org/10.1016/j.jtbi.2006.06.014 PMID: 16890961
48. Lin H, Li QZ. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 2007; 28(9): 1463-6. http://dx.doi.org/10.1002/jcc.20554 PMID: 17330882
49. Lin H, Li QZ. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 2007; 354(2): 548-51. http://dx.doi.org/10.1016/j.bbrc.2007.01.011 PMID: 17239817

50. Li FM, Li QZ. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 2008; 34(1): 119-25. http://dx.doi.org/10.1007/s00726-007-0545-9 PMID: 17514493
51. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 2009; 6: 262-74.http://dx.doi.org/10.2174/157016409789973707
52. Chou KC, Shen HB. Reveiw: Recent advances in developing webservers for predicting protein attributes. Nat Sci 2009; 1(2): 63-92.
53. Chowdhury AS, Reehl SM, Kehn-Hall K, Bishop B, Webb-Robertson BM. Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance. Sci Rep. 2020 Nov 6;10(1):19260. doi: 10.1038/s41598-020-76161-8. PMID: 33159146; PMCID: PMC7648056.

Supplementary Tables

Supplementary Tables 1-3 are not available with this version.

Figures

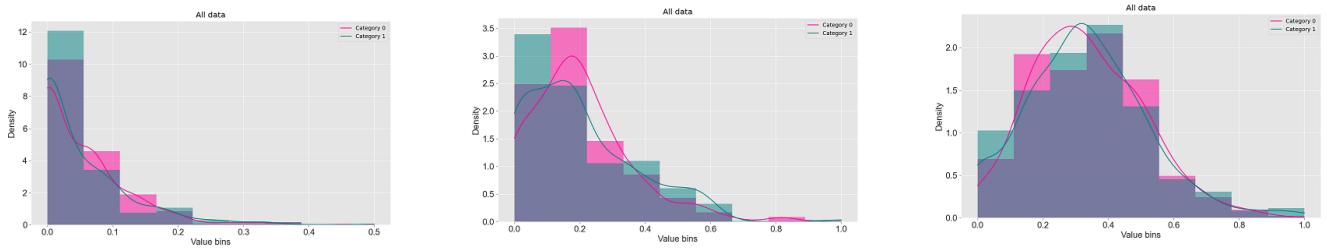

Figure 1

Data distribution for three classes AAC, GAAC and CTDC (from left to right)

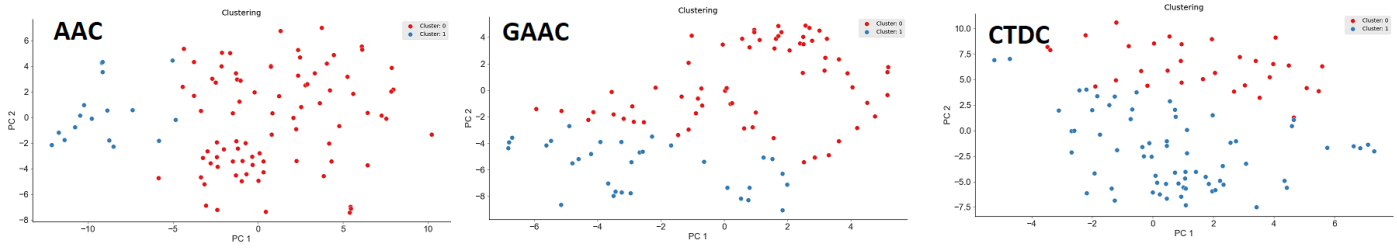

Figure 2

k-means clustering (size=2) for three classes AAC, GAAC and CTDC (from left to right)

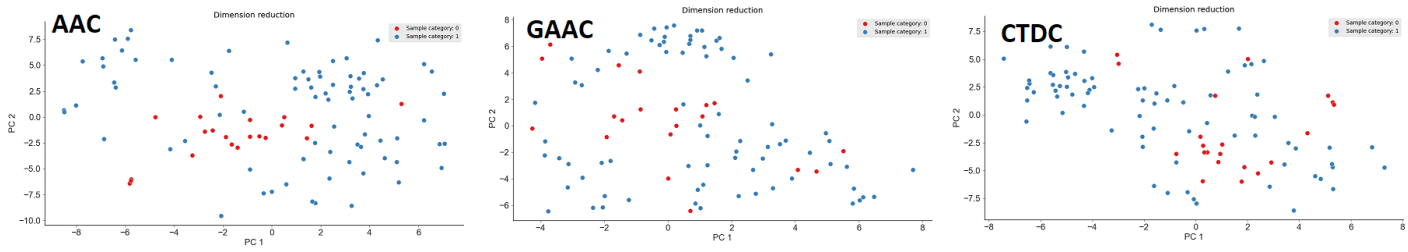

Figure 3

Dimensionality reduction (relevant two principle components) for three classes AAC, GAAC and CTDC (from left to right)

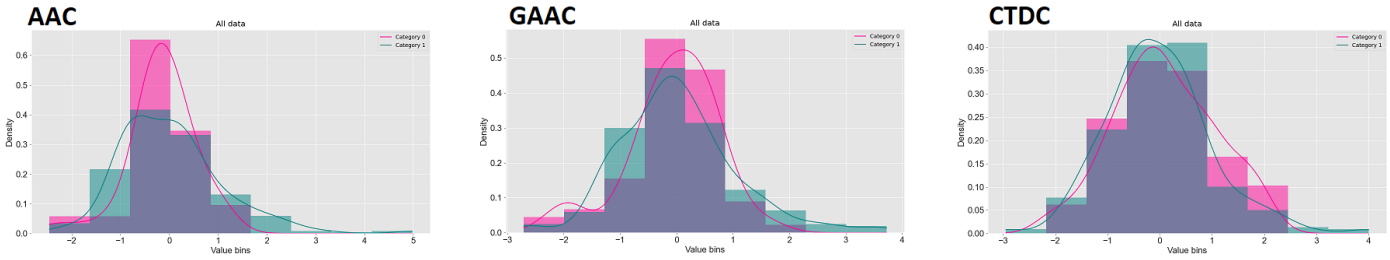

Figure 4

Normalized data (Z-score method chosen) for three classes AAC, GAAC and CTDC (from left to right)

Figure 5

ROC and PRC curve for all the five best model A) GAAC_RF_model B) AAC_k-NN_model C) CTDC_SVM_model D) AAC_SVM E) AAC_RF_model

Figure 6

Box plotof A: AAC_RF_model B: AAC_k-NN_model C: AAC_SVM_model D: GAAC_RF_model E: CTDC_SVM_model

Figure 7

Correlation of models A) k-NN with three descriptors B) RF with three descriptors.