Characterization and Prediction of Dengue Virus Targeting Peptides Based on Combined Amino Acid Composition Descriptors Using Random Forest Algorithm.

Elakkiya Elumalai
Pondicherry University
Suresh Kumar Muthuvel (muthuvels@hotmail.com)
Pondicherry University

Research Article

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

Posted Date: December 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1092942/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. The random forest algorithm was used to predict dengue peptides using grouped amino acid composition, transition and distribution. Here, we have used three descriptors; Amino acid content, Grouped Amino acid composition and Composition, transition and distribution features (CTDC). We have created models and compared with combined model. Using the grouped amino acid composition as input parameters for the random forest algorithm, Our classifier's overall accuracy increased to 88.80%, which was the greatest overall accuracy found in this investigation. 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 peptides.

Highlights

- Amino acid content, grouped amino acid content, CTDC and combined random forest model was developed to predict dengue virus 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 accuracy of the AAC RF and GAAC RF models has improved to 88% and 87%, respectively.

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].

Here, we have proposed a classification algorithm to predict 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. The Random Forest (RF) machine learning algorithm was applied to predict top 5 models for each three descriptors. We compared each model with the combined descriptor model. The descriptors contributing for high model accuracy were, Amino acid content and grouped amino acid composition.

These models were used to predict the dengue virus inhibiting and non-inhibiting peptides. Comparing all developed models, best results were obtained using AAC_RF and GAAC_RF model, this suggests that our classifier is better at predicting dengue viral peptides..

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

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 random forest (RF) as it is more robust algorithm for classification. Here, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions [27]. The normalized dataset (Training set: 102,3; Testing set: 14,3) was taken as input and loaded for machine learning. The random forest algorithm was selected with the following parameters. Tree number: 1000, Number of threads:2, Tree ranges from: 50, Tree ranges to: 500 and Tree steps: 50. The cross validation was set to 5.

2.6 Model validation in testing data

The AAC_RF, GAAC_RF, CTDC_RF and combined_model_RF were validated with testing data. The ROC and PRC curve was ploted. The evaluation metrics was reported.

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

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. We saved these datasets in the Supplementary file (S1).

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 [28, 29]. 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 [30]. 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 plays a vital role in viral defense [31]. 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.

Table 1
Frequent amino acid residues in inhibiting peptides

Feature	Values
G	0.458
F	0.202
W	0.181
N	0.154
A	0.135
I	0.134
D	0.083
L	0.046
E	0.044
V	0.038
K	0.025
P	0.020
T	0.018
C	0.015
R	0.015
Q	0.013
H	0.011
Y	0.007
M	0.001
S	0.000

The alpha helical content in peptides determine its antiviral activity [32]. The data distribution for AAC, GAAC and CTDC is given in Figure 1.

3.3 Machine learning model

The amino acid composition of a protein has been widely utilized for the prediction of peptide categories [33–43]. All descriptors under AAC, GAAC and CTDC was used for clustering (Figure 2) and dimensionality reduction (Figure 3). 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

Figure 4. The normalized data was used as input for model development using Random Forest (RF) algorithm. The RF algorithm is widely used for better understanding and prediction of antiviral peptides [44]. All model (AAC_RF, GAAC_RF, CTDC_RF and combined_RF) metrics was given in Table 2. The ROC and PRC curve for all models are shown in Figure 5. In this table, only the best predictive results of our classifier are illustrated boldly. The boxplot for all models with 8 different evaluation parameters are shown in Figure 6. On looking into the eight parameters, AAC and GAAC models were showing good prediction output. The correlation values between models are given in Figure 7. The highest correlation of 0.9978 was found between AAC_RF and CDTC_RF models.

Table 2
Model metrics for AAC, GAAC, CTDC and combined

Id	Sensitivity	Specificity	Precision	Accuracy	MCC	F1	AUROC	AUPRC
CTDC_RF_model	91.578	41.0	88.422	82.824	0.3357	0.8951	0.8431	0.961
GAAC_RF_model	**92.63**	**61.0**	**92.35**	**87.064**	**0.5235**	**0.9226**	**0.8663**	**0.9683**
AAC_RF_model	**96.842**	**51.0**	**90.55**	**88.802**	**0.5388**	**0.9342**	**0.8487**	**0.9598**
Combined_model	87.368	55.0	91.332	81.956	0.3915	0.886	0.8503	0.9647

The successful predictive performance obtained in our study clearly demonstrated that the combined descriptors (AAC (20 descriptors), GAAC (5 descriptors) and CTDC (39 descriptors)) with Random Forest was quite suitable for predicting these peptides inhibiting dengue virus but overall AAC and GAAC with random forest is the best choice for model development and prediction. The model was evaluated on testing data. The ROC and PRC curve was plotted in Figure 8. The evaluation metrics of all model is given in Table 3.

Table 3
Evaluation metrics of the three model on testing data

Id	Sensitivity	Specificity	Precision	Accuracy	MCC	F1	AUROC	AUPRC
Metrics value_AAC_RL	97.89	95.24	98.94	97.41	0.9147	0.9841	0.9937	0.9986
Metrics value GAAC_RL	97.89	95.24	98.94	97.41	0.9147	0.9841	0.9937	0.9986
Metrics value CTDC_RL	94.74	95.24	98.9	94.83	0.842	0.967	0.990	0.997

Compared to a regular amino acid, the grouped amino acid composition, transition and distribution decreases information redundancy, overfitting and simplifies the protein complexity. To determine which amino acids and biological features were most discriminative between dengue virus inhibiting and non-inhibiting peptides, we analysed differences in amino acids and biological properties. We aimed to create a classifier that could predict dengue virus inhibitory peptides based on the composition, transition, and distribution of grouped amino acids. As a result, these descriptors served as RF's input parameters.

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. An RF algorithm was applied on the three descriptors; AAC, GAAC and CTDC. It was used to predict dengue virus inhibiting peptides. The successful predictive performance obtained in our study clearly demonstrated that these descriptors combined with RF was quite suitable for predicting these two peptide categories. We also developed combined model but the accuracy of this model was comparatively less. The AAC_RF and GAAC_RF model has improved accuracy of 88% and 87% respectively. Based on these data, we believed that our classifier, which uses the scheme of grouped amino acid composition, transition and distribution, 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. 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
28. 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.
29. 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.
30. 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.

31. 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.
32. 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.
33. 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
34. 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
35. 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
36. 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
37. 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
38. 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
39. 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
40. 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
41. 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
42. 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
43. Chou KC, Shen HB. Reveiw: Recent advances in developing webservers for predicting protein attributes. Nat Sci 2009; 1(2): 63-92.
44. 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

Supplementary material is not available with this version

Figures

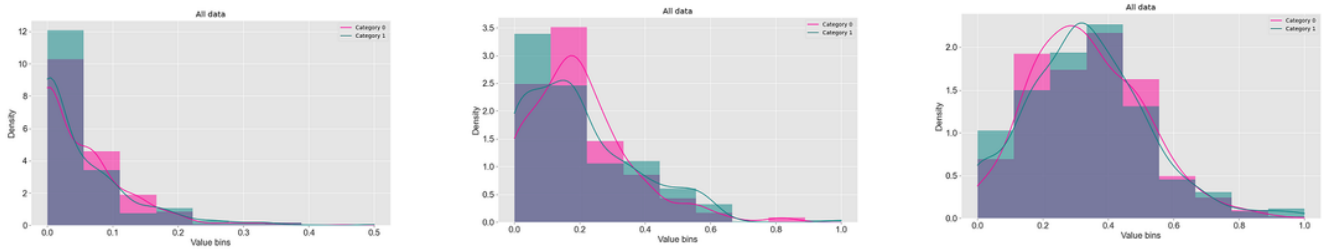

Figure 1

The alpha helical content in peptides determine its antiviral activity [32]. The data distribution for AAC, GAAC and CTDC

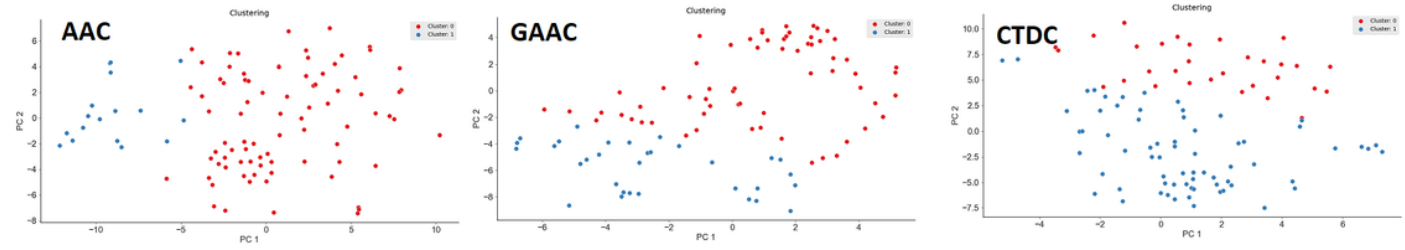

Figure 2

The amino acid composition of a protein has been widely utilized for the prediction of peptide categories [33-43]. All descriptors under AAC, GAAC and CTDC was used for clustering

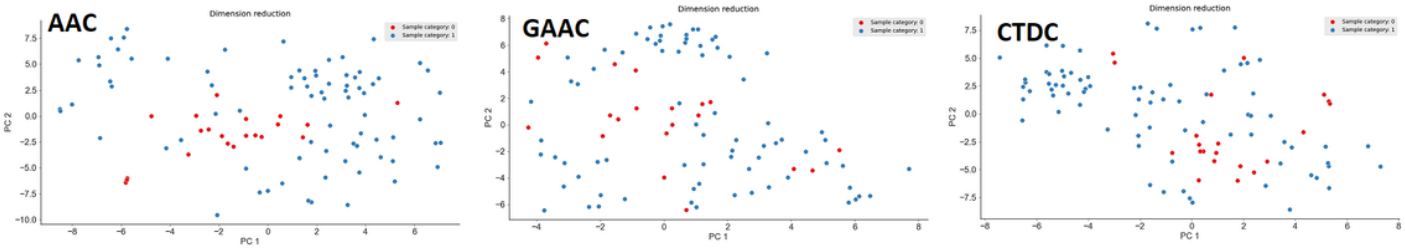

Figure 3

The amino acid composition of a protein has been widely utilized for the prediction of peptide categories [33-43]. All descriptors under AAC, GAAC and CTDC was used for dimensionality reduction

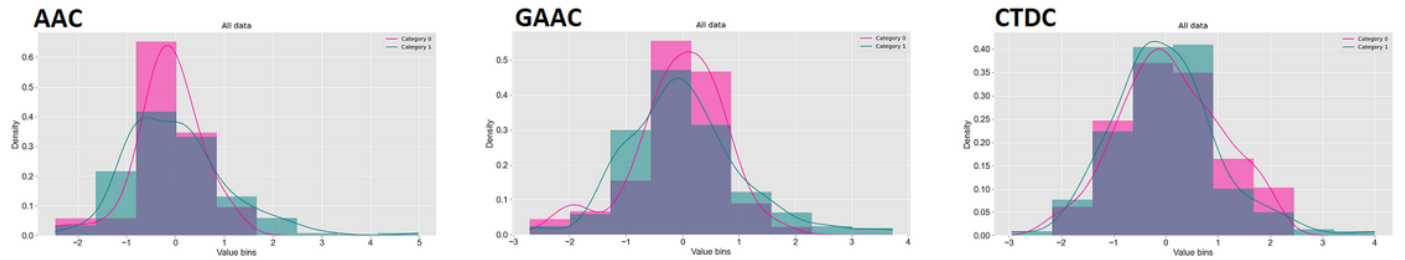

Figure 4

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

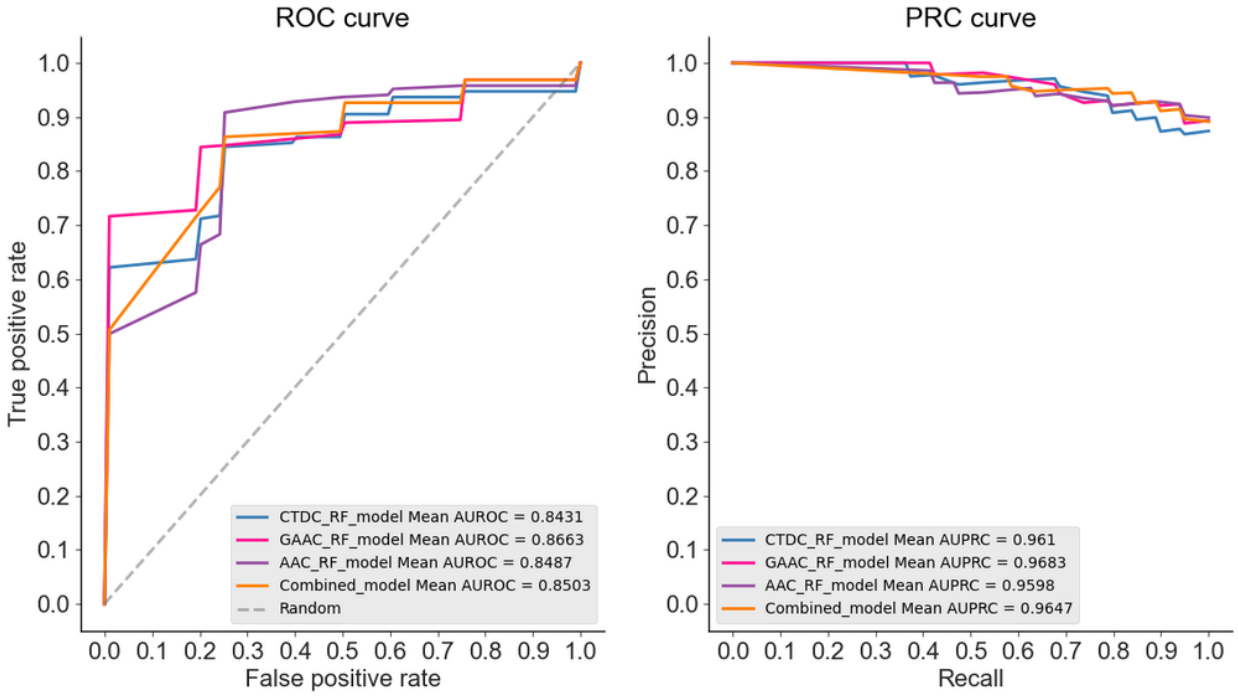

Figure 5

The normalized data was used as input for model development using Random Forest (RF) algorithm. The RF algorithm is widely used for better understanding and prediction of antiviral peptides [44]. All model (AAC_RF, GAAC_RF, CTDC_RF and combined_RF) metrics was given in Table 2. The ROC and PRC curve for all models

Figure 6

In this table, only the best predictive results of our classifier are illustrated boldly. The boxplot for all models with 8 different evaluation parameters

Figure 7

On looking into the eight parameters, AAC and GAAC models were showing good prediction output. The correlation values between models

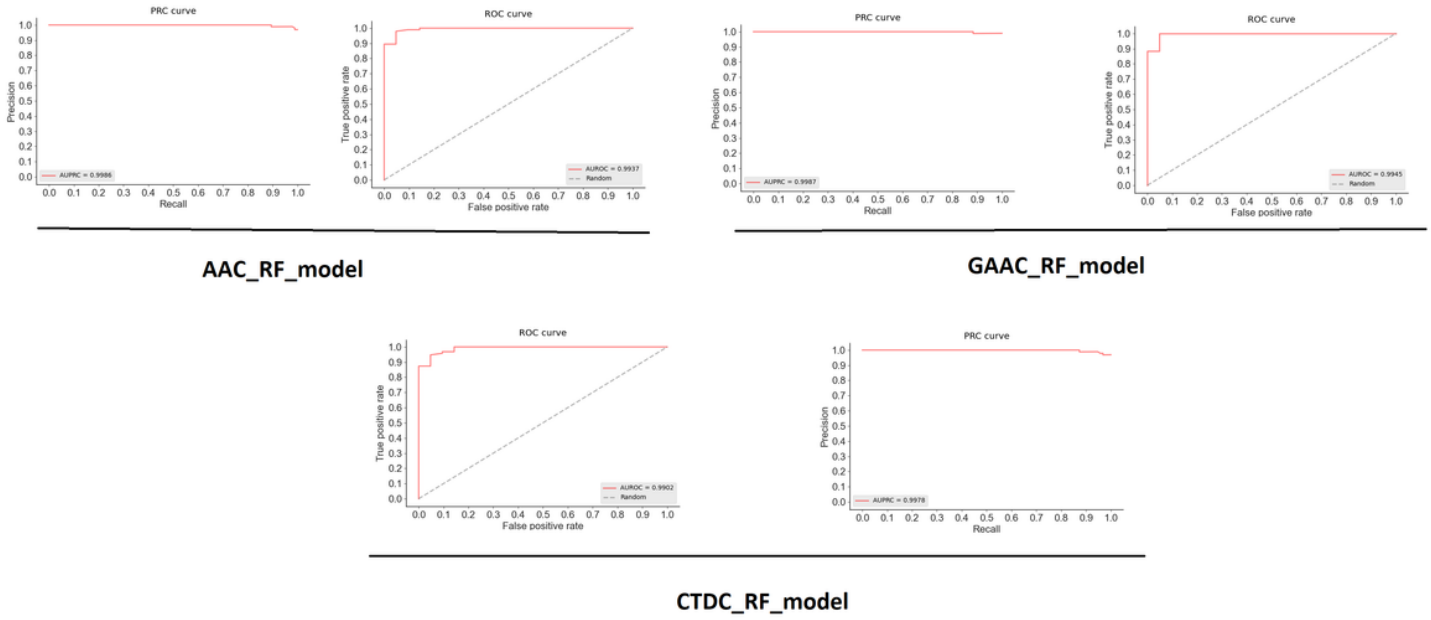

Figure 8

The successful predictive performance obtained in our study clearly demonstrated that the combined descriptors (AAC (20 descriptors), GAAC (5 descriptors) and CTDC (39 descriptors)) with Random Forest was quite suitable for predicting these peptides inhibiting dengue virus but overall AAC and GAAC with random forest is the best choice for model development and prediction. The model was evaluated on testing data. The ROC and PRC curve was plotted