Advances and Applications in Mathematical Sciences
Volume 20, Issue 11, September 2021, Pages 2543-2547

USING RANDOM FOREST TO PREDICT *T*-CELL EPITOPES OF DENGUE VIRUS

SYED NISAR HUSSAIN BUKHARI[1], MUNEER AHMAD DAR[2] and MUJTABA SHAFI[3]

[1,2]National Institute of Electronics
and Information Technology, Srinaga
Jammu and Kashmir, India

[3]University of Kashmir
Srinagar, India
E-mail: nisar@nielit.gov.in
muneer@nielit.gov.in
mujtabashafi@gmail.com

Abstract

The Spread of Dengue virus is considered as one of the severe concern of public health across the world. The cause of an infection is due to the bite of Aedes mosquito. It has been estimated, that around 120 million new cases of dengue infections are reported from about 120 endemic countries. So there is an urgent need to have a vaccine on time for this deadly virus. Here in this study we are trying to predict *T*-cell epitopes of Dengue virus using machine learning which may act as potential candidates for vaccine development. The accuracy of the proposed model has been recorded which is far better than the existing models. The robustness of the proposed model has been validated using repeated *K* fold cross validation technique.

I. Introduction

Dengue virus disease is considered as one of the deadly arbovirus' infections around the world. Around 3.9 billion population reside in areas with a high risk of Dengue virus infection transmission [6, 7]. As per estimates around 390 million individuals are annually infected out of which about 95 million get infection ranging from minimal symptomatic to severe

2010 Mathematics Subject Classification: 60G25.
Keywords: Dengue Virus, Epitopes, Random Forest, Machine Learning, Peptides, *T*-cell Epitopes.
Received October 13, 2020; Accepted November 7, 2020

[8].

The transmission of Dengue virus happens due to the bite of a female mosquito belonging to species Aedes aegypti and genus Aedes [9]. The Dengue virus belongs to Flaviviridae family. There are usually four Dengue virus serotypes (DENV-1, DENV-2, DENV-3 and DENV-4).All are antigenically and genetically related [10].

The viral particles of Dengue virus are enveloped of approximately 50 nanometers in diameter. The infection rate is quite high and there is a need have an effective vaccine on time. The focus of scientists and researchers is to identify the antigenic determinants i.e., epitopes. Epitope is a portion of an antigen (Dengue virus) which is recognized by human immune system (by *T* and *B*-cells of immune system). So recognizing these epitopes is the most important phase in vaccine development, as they play a primary role in activating the human immune system. Number of methods have been proposed to predict these epitopes by various researchers but their accuracy is a concern. Few methods predict peptide binding capacity only [2, 3, 4, 5].

The model that we are proposing shall provide direct prediction of a peptide whether it is a Dengue Virus *T*-Cell epitope or non-epitope with high accuracy.

II. Materials and Methods

We retrieved the experimentally determined *T*-cell epitopes of Dengue virus from IEDB (Immune Epitope Database and Analysis Resource) [1]. The search criteria used for retrieving peptide sequence's is shown in table 1.

Table I. Search criteria used for retrieving peptide sequences.

S. No.	Search Criteria	Value
01	Epitope	Linear epitopes
02	Organism	Dengue Virus(ID:12637)
03	Assays	Positive Assays *T*-cell assays and MHC Ligand assays
04	Host	MHC Restriction Any MHC Restriction

05	Host	Humans
06	Disease	Any Disease

A total of 6380 sequences were retrieved as epitopes. We took sequences of other organisms as non-epitopes. The number of non-epitope sequences used are 2600.

A. Feature Extraction

Feature extraction was done using physicochemical properties of peptide sequences. The following properties were used: Aliphatic index, Molecular weight, Hydrophobicity index, Isoelectric point and Kidera factors. The dataset so generated after feature extraction was used to train the model. The dataset was divided using 70:30 ratios with 70% of the data as training set and 30% as test set. The structure of a dataset is shown in table 2.

Table 2. Structure of dataset of dengue virus *T*-cell epitopes and non-epitopes.

Peptide sequence	Aliphatic index	Molecular weight	Hydrophobicity index	Isoelectric point	Kidera factors	Epitope or non-epitope
YLAGAGLAF	123.7	3.434	-3.33	3.77	1.5	1
ATYGWNLVK	34.345	5.234	-0.464	44.7	2.9	1
APTRVVAAEM	34	4.324	-43.4535	3	0.45	1
TPRMCDDTREEF	543	-4.23	4.345	8	-0.983	0
SVKKDTYLISY	33.2424	4.23	-3.2344	4	1	0
GTSGSPIINR	23.424	43	80.344	1	0.98	1
ATVMDIISRK	23.646	3.44	34	1	4.98	0

B. Machine Learning Models used

We have used random forest algorithm in the current study. Random forest is an ensemble technique and provide more accurate results than using single classifier. It is a collection of decision trees from randomly selected subset of training set. The aggregation of votes from decision trees is used to decide the final class of test data point as shown in Figure 1. The random forest works as:

1. Selection of random samples from dataset.

2. Create decision tree for each randomly selected sample and note down the prediction results from each decision tree.

3. Perform voting for all predicted result.

4. Select prediction result with majority votes as a final prediction.

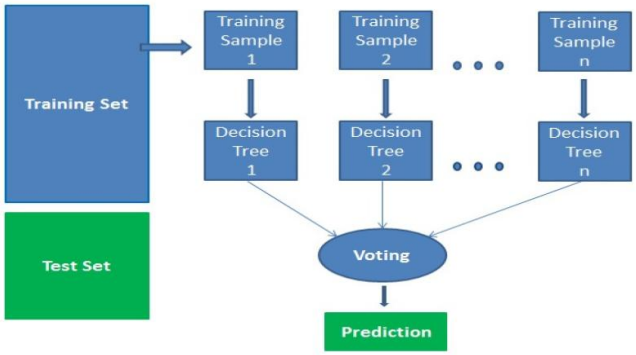

Figure 1. Working of random forest used in the present study.

III. Results

Our model using random forest classifier showed good accuracy than existing methods. The model achieved an accuracy of 86.2% on test set which is overall good and can be treated as a reliable model for predicting *T*-cell epitopes of Dengue virus. In order to check robustness of the model and issues like over fitting and under fitting, a technique call *K*-fold cross validation was used. We created 10 folds of dataset and each fold was executed 5 times and an average accuracy of 85.13% was recorded.

IV. Conclusions

In this paper, model based on random forest has been proposed for prediction of *T*-cell epitopes of Dengue virus. The dataset was obtained from IEDB and later feature extraction was done by taking into account physicochemical properties of peptides. The proposed model efficiently predicts *T*-cell epitopes of Dengue virus in a discrete manner i.e., directly 1 for epitope and 0 for non-epitope. The proposed model achieved an accuracy of 86.2% on the test dataset. The robustness of the proposed model along with

under fitting and over fitting was evaluated using repeated K-fold cross-validation technique. The future work in this area will focus on exploring more machine learning algorithms.

References

[1] P. B. R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler and A. Sette, The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res., 2018. http://www.iedb.org.

[2] L. O. Nielsen M, NN-align, an artificial neural network-based alignment algorithm for MHC class II peptide binding prediction, BMC Bioinformatics 10(1) (2009), 296.

[3] G. J. K. K. Jensen, M. Andreatta, P. Marcatili, S. Buus and N. M. Z. Yan, A. Sette and B. Peters, Improved methods for predicting peptide binding affinity to MHC class II molecules, Immunology 154(3) (2018), 394-406.

[4] C. S. Buus, S. Lauemøller, P. Worning, C. Kesmir, T. Frimurer, B. S. S. A. Fomsgaard, J. Hilden and A. Holm, Sensitive quantitative predictions of peptide-mhc binding by a query by committee artificial neural network approach, Tissue Antigens 62(5) (2003), 378-384.

[5] N. M. M. Andreatta, Gapped sequence alignment using artificial neural networks: application to the MHC class I system, Bioinformatics 32(4) (2015), 511-517.

[6] O. J. Brady, P. W. Gething, S. Bhatt, J. P. Messina, J. S. Brownstein, A. G. Hoen, C. L. Moyes, A. W. Farlow, T. W. Scott and S. I. Hay, Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus. PLoS Negl. Trop. Dis. 2012;6: e1760. doi: 10.1371/journal.pntd.0001760.

[7] R. J. Lin, T. H. Lee and Y. S. Leo Dengue in the elderly: A review, Expert Rev. Anti. Infect. Ther. 15 (2017), 729-735. doi:10.1080/14787210.2017.1358610.

[8] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen and O. Sankoh, et al. The global distribution and burden of dengue. Nature 496 (2013), 504-507. doi: 10.1038/nature12060.

[9] O. J. Brady, M. A. Johansson, C. A. Guerra, S. Bhatt, N. Golding, D. M. Pigott, H. Delatte, M. G. Grech, P. T. Leisnham and R. Maciel-de-Freitas, et al. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. Parasit, Vectors. 2013; 6:351. doi: 10.1186/1756-3305-6-351.

[10] S. B. Halstead and L. F. Dans, Dengue infection and advances in dengue vaccines for children, Lancet Child Adolesc. Heal. 3 (2019), 734-741. doi: 10.1016/S2352-4642(19)30205-6.