Journal of Molecular Evolution (2020) 88:445–452
https://doi.org/10.1007/s00239-020-09943-3

ORIGINAL ARTICLE

FLAVI: An Amino Acid Substitution Model for Flaviviruses

Thu Kim Le[1,2] · Le Sy Vinh[1]

Received: 11 November 2019 / Accepted: 15 April 2020 / Published online: 30 April 2020

Abstract

Amino acid substitution models represent substitution rates among amino acids during the evolution. The models play an important role in analyzing protein sequences, especially inferring phylogenies. The rapid evolution of flaviviruses is expanding the threat in public health. A number of models have been estimated for some viruses, however, they are unable to properly represent amino acid substitution patterns of flaviviruses. In this study, we collected protein sequences from the flavivirus genus to specifically estimate an amino acid substitution model, called FLAVI, for flaviviruses. Experiments showed that the collected dataset was sufficient to estimate a stable model. More importantly, the FLAVI model was remarkably better than other existing models in analyzing flavivirus protein sequences. We recommend researchers to use the FLAVI model when studying protein sequences of flaviviruses or closely related viruses.

Keywords Evolution · Flaviviruses · Dengue virus · West nile virus · Zika virus · Amino acid substitution model

Introduction

An amino acid substitution model is a 20 × 20 matrix describing the substitution rates among 20 amino acids. Amino acid substitution models are essential for investigating the evolutionary relationships among species based on protein sequences. For example, maximum-likelihood phylogenetic tree reconstruction methods require amino acid substitution models for calculating the likelihood values of trees. Distance-based methods use amino acid substitution models to estimate pairwise distances between sequences. Using inappropriate nucleotide/amino acid substitution models would lead to wrong phylogenies (Posada and Crandall 2001). Amino acid substitution models are also crucial for many other protein sequence analyses. For example, amino acid substitution models can be used as score matrices for protein sequence similarity search or protein sequence alignment. The roles and applications of the amino acid substitution models are summarized by (Thorne 2000).

A general time-reversible nucleotide substitution model consists of only 8 free parameters that can be easily estimated from an alignment under the study. An amino acid substitution model contains 208 parameters, therefore, individual alignments do not provide enough information for correctly estimating such large number of parameters. Thus, amino acid substitution models must be estimated from large protein datasets in advance. General amino acid substitution models such as LG (Le and Gascuel 2008) have been estimated from multiple alignments including various species and suitable for analyzing general protein alignments.

Viruses have a short generation time and a large population size, therefore, they can evolve quickly to adapt environmental changes or immune responses from hosts. That results in various amino acid substitution patterns during the rapid evolution of viruses. A number of virus-specific amino acid substitution models have been estimated such as rtREV for retroviruses (Dimmic et al. 2002), HIVb and HIVw for HIV viruses (Nickle et al. 2007), FLU for influenza viruses (Cuong et al. 2010). Experiments showed that the virus-specific models were better than general models when analyzing protein sequences from their corresponding viruses. For example, the FLU model is much better than other models in analyzing protein sequences from influenza viruses (Cuong et al. 2010).

Handling Editor: **Keith Crandall**.

✉ Le Sy Vinh
vinhls@vnu.edu.vn

[1] University of Engineering and Technology, Vietnam National University Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi 10000, Vietnam

[2] Hanoi University of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi 10000, Vietnam

Recently, flaviviruses have remerged and caused life-threatening outbreaks, especially in tropical and subtropical regions (Daep, Muñoz-Jordán, and Eugenin 2014; Hatcher et al. 2017). Hence, estimating an amino acid substitution model for flaviviruses is imperative to properly characterize the evolution of the viruses. We note that protein sequences are appropriate for studying viruses because the rapid evolution of viruses might make nucleotide sequences saturated. In this study, we collected available protein sequences from the flavivirus genus and employed the maximum-likelihood method to estimate an amino acid substitution model, called FLAVI. Experiments showed that FLAVI was robust and better than existing models in analyzing protein sequences from flaviviruses. The FLAVI model will enhance evolutionary studies of flaviviruses.

Material and Methods

Material

Flavivirus genus includes West Nile, Dengue, Zika, and some other viruses. They are small enveloped viruses whose complete genomes consist of from 9,500 to 12,500 nucleotides. The RNA genome of these viruses encodes three structural proteins (i.e., E, PrM, and C) and seven non-structural proteins (i.e., NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5) (Bollati et al. 2010).

The protein sequences of West Nile, Dengue, and Zika viruses are available from Virus Variation Resource at NCBI (https://www.ncbi.nlm.nih.gov/genomes/VirusVariation/) (Hatcher et al. 2017). We downloaded all protein sequences (available up to April 2019) from the three viruses to form a dataset **D** of 11,392 distinct sequences (i.e., 603 Zika sequences, 2091 West Nile sequences, and 8698 Dengue sequences).

The protein sequences in **D** were randomly divided into two equal parts: one for training the model and the other for testing the model. All sequences from the same virus species and protein type in the training dataset were aligned together using the MUSCLE program (Edgar 2004) to create a multiple sequence alignment. As there were three virus species and 10 protein types, the training dataset consisted of 30 training multiple sequence alignments. Similarly, the testing dataset included 30 testing alignments each corresponding to one protein type of a virus species.

Methods

Substitutions among amino acids during the evolution are modeled by a time-homogeneous, time-continuous, and time-reversible Markov process and described by a 20 × 20 instantaneous substitution rate matrix $Q = \{q_{xy}\}$ where q_{xy} represents the number of substitutions between two different amino acids x and y per a time unit (q_{xx} is assigned such that the sum of all elements on row x of matrix Q equals to zero). Since the amino acid substitution process is assumed to be time-reversible, the matrix Q can be decomposed into a symmetric exchangeability rate matrix $R = \{r_{xy}\}$ and an amino acid equilibrium frequency vector $\Pi = \{\pi_x\}$. Technically, if $x \neq y, q_{xy} = \pi_y r_{xy}$, otherwise, $q_{xx} = -\sum_y q_{xy}$. Note that in phylogenetic tree construction, the branch lengths normally reflect the number of mutations, thus, the matrix Q is normalized as follows:

$$Q = \frac{Q}{\mu} \text{ where} \mu = -\sum q_{xx}$$

Given a dataset $\mathbf{D} = (D_1, \dots, D_n)$ of n multiple amino acid sequence alignments, let $\mathbf{T} = (T_1, \dots T_n)$ be the tree set corresponding to the dataset $\mathbf{D}$, i.e., T_i is the tree of alignment D_i. The maximum-likelihood estimation method determines the tree set $\mathbf{T}$ and a model M to maximize the likelihood value $L(M, \mathbf{T};\mathbf{D})$. We assume that the amino acid substitutions among alignments and sites are independent, thus, the likelihood value $L(M, \mathbf{T};\mathbf{D})$ can be calculated as follows:

$$L(M, \mathbf{T};\mathbf{D}) = \prod_{i=1}^{n} L(M, T_{\mathrm{i}};D_i) = \prod_{i=1}^{n} \prod_{j=1}^{l_i} L\left(M, T_{\mathrm{i}};D_{ij}\right)$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{l_i} P(D_{ij}|M, T_{\mathrm{i}})$$

where l_i is the length of alignment D_i; and D_{ij} is the data at site j of alignment D_i. The likelihood value $L(M, T_{\mathrm{i}}; D_{ij})$ can be calculated by the conditional probability $P(D_{ij}|M, T_{\mathrm{i}})$ of data D_{ij} given the model M and the tree T_i.

It is well-known that amino acid substitution rates vary among sites. We can properly incorporate the site rate heterogeneity by determining site rate models $V = (V_1, \dots, V_n)$ for alignments $\mathbf{D}$, i.e., V_i is the site rate model of alignment D_i. Typically, a site rate model combines a gamma distribution model and an invariant rate model (Yang 1993). The likelihood value $L(M, \mathbf{T}, \mathbf{V}; \mathbf{D})$ is now technically calculated as follows:

$$L(M, \mathbf{T}, \mathbf{V};\mathbf{D}) = \prod_{i=1}^{n} \prod_{j=1}^{l_i} L\left(M, T_{\mathrm{i}}, V_i;D_{ij}\right) = \prod_{i=1}^{n} \prod_{j=1}^{l_i} P\left(D_{ij}|M, T_{\mathrm{i}}, V_i\right)$$

where $P\left(D_{ij}\middle|M, T_i, V_i\right)$ is the conditional probability of data D_{ij} given the model M, the tree T_i, and the site rate model V_i.

Estimating the parameters of a model M is computationally difficult because we have to simultaneously estimate the trees $\mathbf{T}$, the site rate models $\mathbf{V}$ and the amino acid substitution model M. A number of approximate maximum-likelihood methods have been proposed to estimate parameters of a model M from large datasets (Whelan and Goldman 2001; Dang et al. 2011; Le and Gascuel 2008). The methods showed that the parameters of model M can be accurately estimated using nearly optimal trees **T** and site rate models **V**. Thus, we can iteratively estimate the trees **T**, site rate models **V** and model M.

The most time-consuming step in the estimation process is determining maximum-likelihood trees from large alignments. To overcome the obstacle, alignment-splitting algorithms have been proposed to divide large alignments into smaller alignments such that smaller alignments still contain sufficient phylogenetic information to estimate the model M while significantly reduce the time for constructing maximum-likelihood trees (Dang et al. 2014).

The approximate maximum-likelihood procedure to estimate parameters of a model M from a set of protein alignments **D** is composed of three main steps (i.e., alignment splitting, tree construction, and model estimation) and illustrated in Fig. 1. The estimation procedure is described as following:

- *Alignment splitting step* The step divides large alignments in the training dataset **D** into smaller sub-alignments to avoid computational burden in building maximum likelihood trees. In this paper, we employed the tree-based splitting algorithm (Dang et al. 2014) to divide large training alignments into smaller sub-alignments each containing from 25 to 50 sequences. We obtained 150 sub-alignments for estimating the parameters of M.
- *Tree construction step* The step builds maximum-likelihood trees for sub-alignments. For each sub-alignment D_i, we determined the best-fit models (i.e., the amino acid substitution model and the site rate model), and

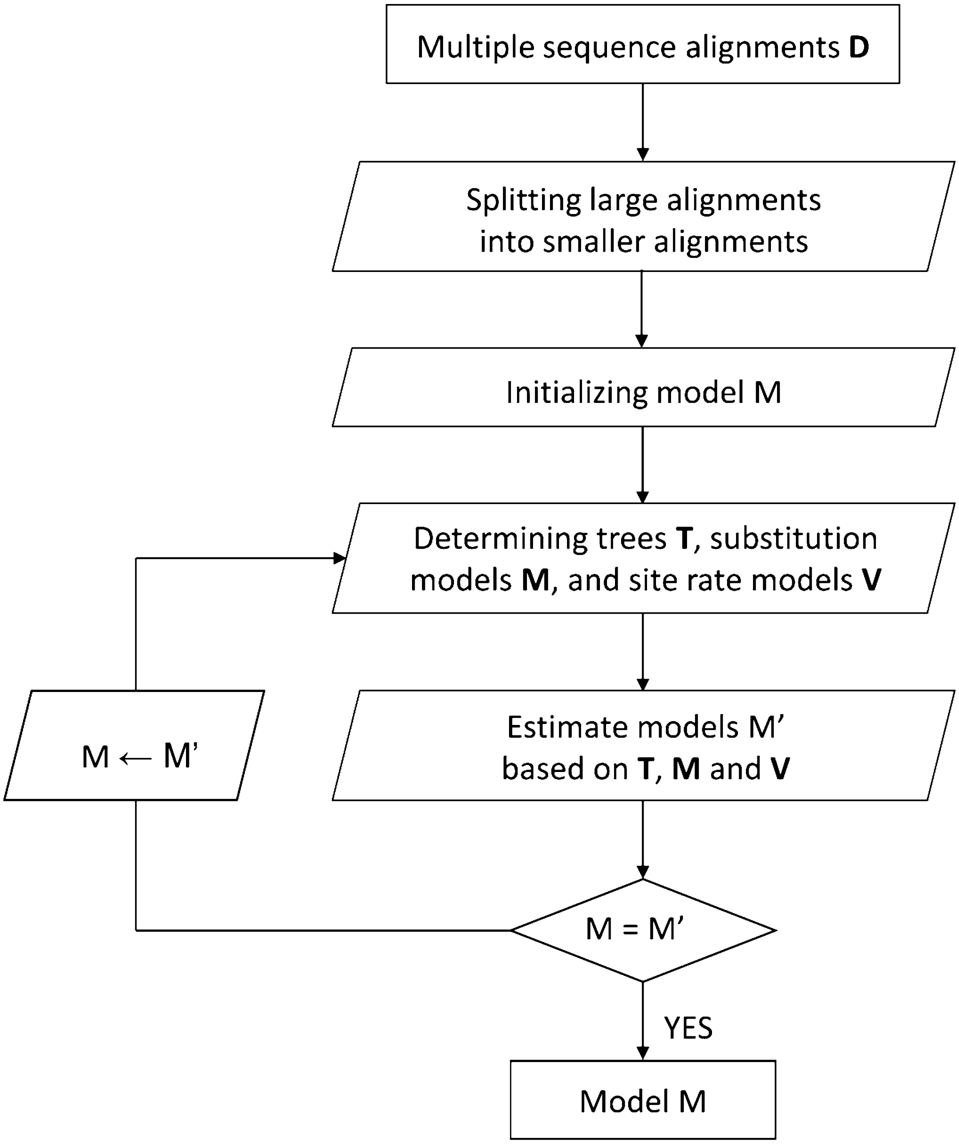
Multiple sequence alignments D
Splitting large alignments into smaller alignments
Initializing model M
Determining trees T, substitution models M, and site rate models V
Estimate models M' based on T, M and V
M = M'
YES
Model M
M ← M'

Fig. 1 An approximate maximum-likelihood procedure to estimate an amino acid substitution model from a set of protein alignments

consequently constructed the maximum-likelihood tree for D_i. The best-fit amino acid substitution model for D_i was selected from a set of models including the current model *M* and other existing models (i.e., FLU, HIVb, HIVw, JTT, WAG, LG) using the ModelFinder program (Kalyaanamoorthy et al. 2017). The maximum likelihood trees were constructed by the IQ-TREE program (Nguyen et al. 2015).

- *Model estimation step* The step estimates parameters of a new model *M’* from the sub-alignments based on the trees and best-fit models obtained from the tree construction step. The parameters of *M’* were estimated using QMaker (Bui Quang et al. 2020), our newly developed function in the IQ-TREE package.

  If the new model *M’* is highly similar to the current model *M* (i.e., the correlation greater than 0.999), the estimation procedure stops and considers the model *M’* as the final model. Otherwise, it assigns *M* by *M’* and performs the tree construction and model estimation steps again. Normally, the estimation procedure stops after 3 iterations (Vinh et al. 2017; Dang et al. 2014; Cuong et al. 2010).

The script to estimate parameters of a model *M* from a training dataset is available at https://github.com/thulekm/flavi.

An amino acid substitution model consists of 208 free parameters, i.e., 189 parameters from the matrix *R*, and 19 parameters from the vector π, therefore, it is typically estimated from large datasets. We assessed that if the collected flavivirus protein dataset was sufficient to estimate a stable model. To this end, we randomly divided the dataset **D** into two parts: the first part **D1** used to estimate the model FLAVI-1, and the second part **D2** used to estimate the model FLAVI-2. The correlation between FLAVI-1 and FLAVI-2 models indicates the stability of FLAVI models.

We compared FLAVI with seven other models including three popular general models, i.e., JTT (Jones et al.1992), LG (Le and Gascuel 2008), and LG4X (Le et al. 2012), and four virus-specific models, i.e., FLU for influenza viruses (Cuong et al. 2010), HIVb/HIVw for HIV viruses (Nickle et al. 2007), and rtREV for retroviruses (Dimmic et al. 2002) on the testing dataset. Note that the LG4X is a mixture model that includes four different substitution rate matrices corresponding to four different site rate categories. For each testing alignment, we used IQ-TREE to construct eight maximum-likelihood trees corresponding to eight models (the gamma distribution model with 4 categories and the invariant rate model were employed for the site rate heterogeneity). We compared the performance of different models using the Akaike information criterion (AIC) (Hirotugu 1974), i.e., the smaller AIC score indicates the better model.

We also applied the approximately unbiased test (AU test) of phylogenetic tree selection (Hidetoshi Shimodaira 2002) to examine if the tree constructed with the best model was significantly better than trees constructed with other models. Technically, given a testing alignment, we constructed maximum likelihood trees with different models, and subsequently used the CONSEL program (H. Shimodaira and Hasegawa 2001) for assessing the confidence levels of the models.

Results and Discussions

Model Analysis

FLAVI-1 and FLAVI-2 models were estimated from **D1** and **D2** datasets, respectively. The Pearson correlation between FLAVI-1 and FLAVI-2 models was over 0.99 (i.e., the correlation between two exchangeability rate matrices was 0.994; the correlation between two amino acid frequency vectors was 0.991). The high correlation between FLAVI-1 and FLAVI-2 models affirmed that the collected dataset was sufficient to estimate a stable model for flaviviruses.

Table 1 shows the correlations among models. Generally, the existing models are not highly correlated (e.g., the correlation between LG and HIVb is only 0.8), even between virus-specific models (e.g., the correlation between HIVb and FLU is 0.86). The correlations between FLAVI and existing models ranges from 0.67 (rtREV) to 0.92 (FLU) indicating that FLAVI is not highly correlated with the existing models. We note a very low correlation between FLAVI and rtREV that might be due to the lack of retrovirus protein

Table 1 The Pearson correlations between FLAVI and other models. The values in the top triangle represent the correlations between exchangeability matrices, while values in the low triangle are the correlations between frequency vectors

	FLAVI	JTT	LG	HIVw	HIVb	FLU	rtREV
FLAVI	–	0.88	0.77	0.89	0.91	0.92	0.67
JTT	0.90	–	0.91	0.80	0.90	0.87	0.83
LG	0.85	0.96	–	0.65	0.80	0.81	0.95
HIVw	0.59	0.58	0.55	–	0.85	0.84	0.54
HIVb	0.86	0.88	0.89	0.66	–	0.86	0.71
FLU	0.73	0.80	0.72	0.84	0.72	–	0.75
rtREV	0.80	0.86	0.87	0.59	0.90	0.67	–

sequences (i.e., only 33 sequences) used for estimating the rtREV model. As anticipated, the correlations between FLAVI with general models are remarkably low (i.e., 0.76 between FLAVI and LG or 0.88 between FLAVI and JTT) as the general models were estimated from diverse datasets and unable to properly reflect the rapid changes of flaviviruses. As the mixture model LG4X consisted of four matrices, we did not measure the correlation between FLAVI and LG4X.

The amino acid frequencies and exchangeability rates of FLAVI, HIVb, and LG are illustrated in Figs. 2 and 3, respectively. We observe a number of notable differences between these models. For example, the frequency of M (methionine) in FLAVI is about twice higher than that in HIVb; or the substitution rate between amino acid A (alanine) and amino acid V (valine) in the FLAVI is much higher than those in HIVb and LG. Figure 4 illustrates a deeper look at the relationships between these models. A large number of big circles in the figure indicate the high discrepancy between models. For example, nearly half of the entries in the FLAVI matrix are at least five times smaller or greater than the corresponding elements in the HIVb matrix.

Table 2 summarizes comparisons between FLAVI, HIVb, and LG models. The results confirm that the existing models are remarkably different from the FLAVI model, thus, they are unable to properly represent the amino acid substitution process of flaviviruses.

Model Performance

We compared the performance of FLAVI and other models in building maximum-likelihood trees from the testing dataset, i.e., we assessed whether FLAVI enables us to produce trees with better AIC score. We also examined the difference between tree topologies constructed by FLAVI and other models. Specifically, for each model we computed the average AIC score of trees constructed from the testing dataset (see Table 3). On average, FLAVI produced better trees than other models. HIVb and FLU were the second and third best models, respectively. As anticipated, the virus-specific models except rtREV were better than general models in building maximum-likelihood trees for flaviviruses. The rtREV model was worse than other models as it was estimated from an insufficient dataset. The poor performance of rtREV warns that amino acid substitution models must be estimated from reasonably large datasets.

We also compared the performance of FLAVI and other models at the individual alignment level. The FLAVI model found the best trees for 28 out of 30 testing alignments and

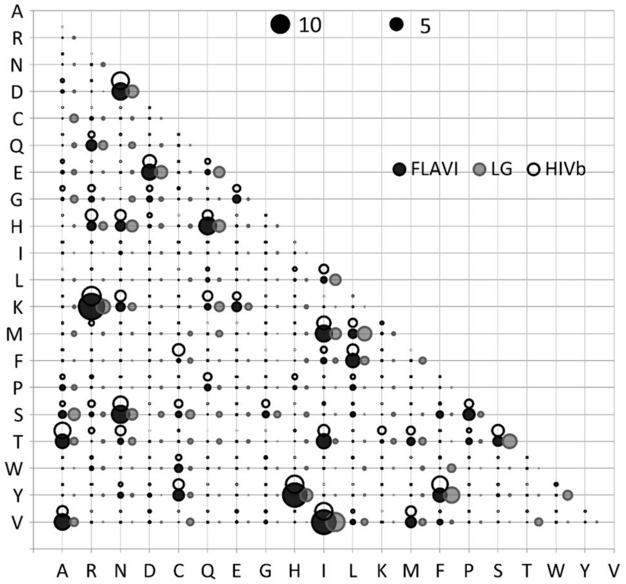

Fig. 3 The exchangeability rates in FLAVI, HIVb and LG models. The black, grey and white circles at row *X* and column *Y* represent the exchangeability rates between amino acids *X* and *Y* in FLAVI, LG, and HIVb models, respectively

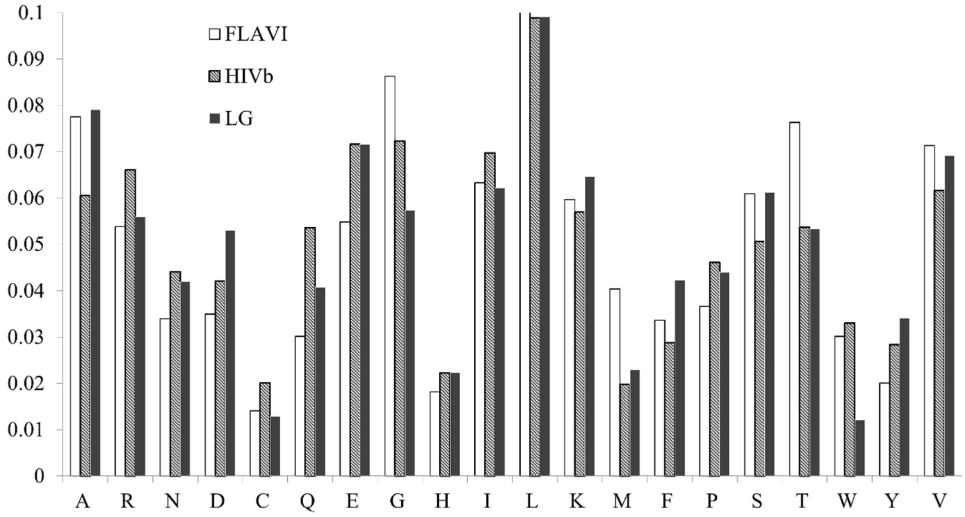

Fig. 2 The amino acid frequencies of FLAVI, LG and HIVb models

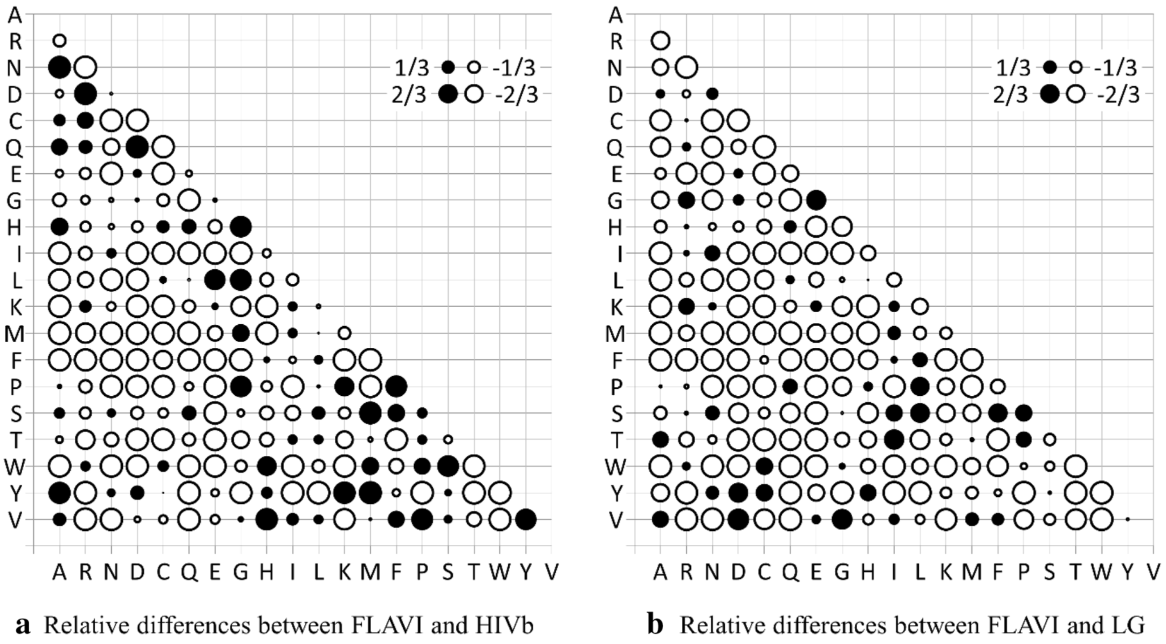

Fig. 4 The left figure is the relative differences between exchangeability coefficients in FLAVI and HIVb. Each bubble represents the value of $\left(FLAVI_{XY} - HIVb_{XY}\right)/\left(FLAVI_{XY} + HIVb_{XY}\right)$. Values 1/3 and 2/3 mean that the coefficient in FLAVI is two and five times larger than that in HIVb, respectively. Values $-1/3$ and $-2/3$ mean that the coefficient in FLAVI is two and five times smaller than that in HIVb, respectively. Similar explanations for the right figure but between FLAVI and LG

Table 2 Relative differences between FLAVI vs HIVb and FLAVI vs LG models. The number at row "Twice" and column "FLAVI > HIVb" is the number of coefficients in FLAVI that are at least twice larger than that in HIVb. Similar explanations for other numbers in the table

	FLAVI > HIVb	HIVb > FLAVI	FLAVI > LG	LG > FLAVI
Twice	31	96	24	120
Five	18	71	7	92

Table 3 The comparisons between FLAVI and other models in building maximum-likelihood trees on 30 testing alignments. The smaller AIC value, the better the model

	AIC per alignment	AIC per site
FLAVI	6649	34.8
HIVb	6722	35.2
FLU	6745	35.3
JTT	6771	35.5
HIVw	6799	35.6
LG4X	6861	35.9
LG	6889	36.1
rtREV	6964	36.5

the second-best trees for the two remaining alignments. The HIVb model was the second-best model for building maximum-likelihood trees, i.e., finding the best trees for 2 out of 30 alignments, and the second-best trees for 18 other alignments.

Figure 5 shows the comparison between FLAVI and HIVb (i.e., the second-best model). The FLAVI model was better than the HIVb model for 28 testing alignments. The AU tests showed that trees constructed with FLAVI were significantly better (i.e., 0.95 confidence level) than trees constructed with HIVb for 16 cases. There were 12 cases that trees constructed with FLAVI had higher likelihood values, however, they were not significantly better than trees constructed with HIVb. Analyzing the trees revealed that they have polytomy-like topologies, i.e., external branches connecting leaves are much longer than internal branches.

Although the number of protein sequences from Zika, West Nile and Dengue viruses in the training dataset were unbalanced (i.e., 5.3% Zika sequences, 18.3% West Nile sequences, and 76.4% Dengue sequences), the FLAVI model performed well on testing alignments from all three virus kinds. Specifically, FLAVI outperformed the other models for all Dengue and West Nile testing alignments. It was

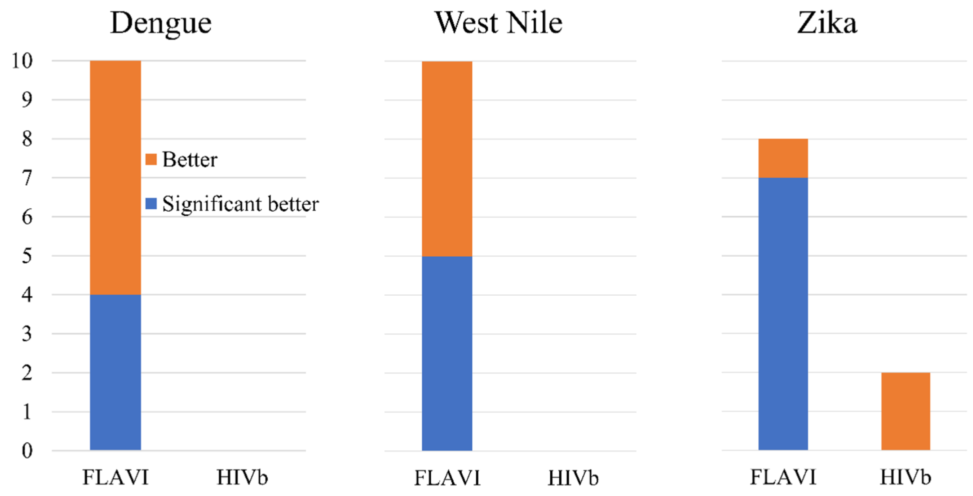

Fig. 5 The comparison between FLAVI and HIVb models in building maximum-likelihood trees on Dengue, West Nile and Zika testing alignments. Better: the tree has higher likelihood value. Significant better: the tree has significant higher likelihood value using the approximately unbiased test

the best model for 8 out of 10 Zika tests. In addition, we estimated the Dengue model using only Dengue training alignments, and subsequently measured its performance on West Nile and Zika testing alignments. The Dengue model outperformed the other existing models (i.e., FLU, HIVb, HIVw, rtREV, JTT, LG, LG4X) on 18 out of 20 West Nile and Zika tests. It was the second-best model for the two remaining tests. From the results, it can be extrapolated that the FLAVI model will perform well not only for three kinds of flaviviruses in the training dataset but also for other flavivirus types.

Conclusions

Flavivirus genus, including emerging viruses such as West Nile, Dengue, or Zika, are causing an expanding threat in the public health. We collected available protein sequences of flaviviruses to estimate the FLAVI amino acid substitution model. Experiments confirmed that the collected dataset was sufficient to estimate a stable model. The FLAVI model is remarkably different from the existing models including general models and virus-specific models. Thus, the FLAVI model should be used to properly describe the amino acid substitution process of flaviviruses.

Experiments showed that FLAVI was better than other models in analyzing flavivirus protein sequences. FLAVI helped produce better phylogenetic trees than the existing models in almost all testing alignments indicating the fitness of FLAVI for flavivirus protein sequences. The FLAVI model should be used as the default model for analyzing protein sequences from flaviviruses. As FLAVI might not be the best-fit model for some cases, researchers might use model selection programs such as ModelFinder to determine the best-fit model for alignments under the study.

The discrepancy between trees constructed with FLAVI and other models was considerably high. The average normalized Robinson–Foulds distance (Robinson and Foulds 1981) between trees constructed with FLAVI and other models ranged from 0.61 to 0.66 (e.g., the average Robinson–Foulds distance between FLAVI-based trees and HIVb-based trees was 0.63). The high discrepancy between trees constructed with different models implied that trees constructed with the existing models contained a considerable number of incorrect clades (Vinh et al. 2017). Examining two trees constructed from all three Dengue, West Nile and Zika virus types based on the FLAVI and HIVb models revealed that both trees were able to correctly classify sequences from three virus types into three distinct clades. The large Robinson–Foulds distance between the trees was due to short branches inside the clades (i.e., the average branch length was about 0.006). As the existing models were not estimated from flavivirus data, they were unable to correctly resolve the relationships among closely related sequences from the same virus type.

The total length of trees with FLAVI (i.e., 63.4) was longer than that of trees with the other models (i.e., ranged from 55.6 to 61.3). The results indicated that FLAVI helped properly model the rapid evolution of flaviviruses. The flavivirus trees might contain polytomy-like structures, therefore, we should use standard nonparametric bootstrap method (Felsenstein 1985) or fast approximation bootstrap method (Hoang et al. 2017) to assess the reliability of clades in the maximum-likelihood constructed trees.

Acknowledgements This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant number 102.01.2019.06.

Author Contributions VLS designed the study and experiments. LKT performed experiments. Both authors analyzed experimental results, wrote and approved the final manuscript.

Data Availability The FLAVI data is available at https://github.com/thulekm/flavi

Compliance with Ethical Standards

Conflict of interest The authors declare no competing interests.

References

Bollati M, Alvarez K, Assenberg R, Baronti C, Canard B, Cook S, Coutard B et al (2010) Structure and functionality in flavivirus NS-proteins: perspectives for drug design. Antiviral Res 87(2):125–128. https://doi.org/10.1016/j.antiviral.2009.11.009

Minh BQ, CD Cao, VL Sy, R Lanfear (2020) QMaker: estimating empirical models of protein evolution from large collections of alignments. Submitted

Cuong D, Quang Le, Olivier G, Vinh LS (2010) FLU, an amino acid substitution model for influenza proteins. BMC Evol Biol 10:99

Daep CA, Muñoz-Jordán JL, Eugenin EA (2014) Flaviviruses, an expanding threat in public health: focus on dengue, West Nile, and Japanese encephalitis virus. J Neurovirol 20(6):539–560. https://doi.org/10.1007/s13365-014-0285-z

Dang CC, Lefort V, Vinh LS, Le QS, Gascuel O (2011) Replacementmatrix: a web server for maximum-likelihood estimation of amino acid replacement rate matrices. Bioinformatics 27(19):2758–2760. https://doi.org/10.1093/bioinformatics/btr435

Dang CC, Vinh LS, Gascuel O, Hazes B, Le QS (2014) FastMG: a simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets. BMC Bioinform 15(1):341. https://doi.org/10.1186/1471-2105-15-341

Dimmic MW, Rest JS, Mindell DP, Goldstein RA (2002) RtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol 55(1):65–73. https://doi.org/10.1007/s00239-001-2304-y

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797. https://doi.org/10.1093/nar/gkh340

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39(4):783–791. https://doi.org/10.2307/2408678

Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, Schaffer AA, Rodney Brister J (2017) Virus variation resource-improved response to emergent viral outbreaks. Nucleic Acids Res 45(D1):D482–D490. https://doi.org/10.1093/nar/gkw1065

Hirotugu A (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723. https://doi.org/10.1109/TAC.1974.1100705

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2017) UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35(2):518–522. https://doi.org/10.1093/molbev/msx281

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8:275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14(587):589. https://doi.org/10.1038/nmeth.4285

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25(7):1307–1320. https://doi.org/10.1093/molbev/msn067

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268–274. https://doi.org/10.1093/molbev/msu300

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, SL, Pond (2007) HIV-specific probabilistic models of protein evolution. PLoS ONE 2(6):e503. https://doi.org/10.1371/journal.pone.0000503

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA 98:13757–13762. https://doi.org/10.1073/pnas.241370698

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247. https://doi.org/10.1093/bioinformatics/17.12.1246

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Syst Biol 51(3):492–508. https://doi.org/10.1080/10635150290069913

Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol 29:2921–2936

Thorne JL (2000) Models of protein sequence evolution and their applications. Curr Opin Genet Dev 10:602–605. https://doi.org/10.1016/S0959-437X(00)00142-8

Vinh LS, Dang CC, Le QS (2017) Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. BMC Evol Biol 17:136. https://doi.org/10.1186/s12862-017-0987-y

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18(5):691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10(6):1396–1401