2018 10th International Conference on Knowledge and Systems Engineering (KSE)

Building a Specific Amino Acid Substitution Model for Dengue Viruses

Thu Le Kim
Hanoi University of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Ha Noi
thu.lekim@hust.edu.vn

Cuong Dang Cao
VNU-University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Ha Noi
cuongdc@vnu.edu.vn

Vinh Sy Le
VNU-University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Ha Noi
vinhls@vnu.edu.vn

***Abstract*—Phylogenetic trees inferred from protein sequences are strongly affected by amino acid substitution models. Although different amino acid substitution models have been proposed, only a few were estimated for specific species such as the FLU model for influenza viruses. Among the most dangerous viruses for human health, dengue is always on top and the cause of dengue fever up to 100 million people per year. In this study, we built a specific amino acid substitution model for dengue protein sequences, called DEN. The dengue protein sequences were obtained from the NCBI dengue database and the model was estimated using the maximum likelihood method. Experiments showed that the new model DEN helped to build better phylogenetic trees than other existing models. We strongly recommend researchers to use the DEN model for analyzing dengue protein data.**

Keywords— dengue virus, amino acid substitution model, phylogeny tree.

I. BACKGROUND

Amino acid substitution models (models for short) have an important role in many protein analyses such as measuring the genetic distance among protein sequences or building phylogenetic trees. The standard amino acid substitution model consists of two components: a 20 x 20 instantaneous substitution rate matrix, in which element at x^{th} row and y^{th} column (except entries on the main diagonal) represents the substitution rate between amino acid x and y per time unit, and a vector of 20 amino acid frequencies [1][2].

There are two main approaches to estimate amino acid substitution models, the distance-based approach and the maximum likelihood approach. The distance-based approach assumes that the exchangeability probability from amino acid x to y over a period is linear to the number of amino acid substitutions between x and y. Thus, the rates are directly estimated from data. The two most widely used models deriving from the distance-based method are PAM and JTT [3][4]. The advantage of this approach is the fast estimation time, however, it is only applicable to closely related amino acid sequences (i.e., the similarity among sequences is greater than or equal to 85%). The maximum likelihood method was proposed by Felsenstein [1] with the goal to estimate both phylogenetic trees and amino acid models together in order to maximize the likelihood of data [1][5]. Though, the estimation process is much more time-consuming, but it does not constrain sequences to be highly similar. The models estimated from general protein sequences are usually called general models. LG is one of those general models as it was estimated from general protein sequences in the Pfam database [2]. Currently, LG is usually considered as the best general amino acid substitution model.

Although a number of general models have been calculated from diverse databases, the evolutionary processes of different species vary considerably. As a result, the general models might not fit for a specific species. A number of models have been built for different viruses. Among dangerous viruses, HIV viruses, retrovirus and influenza viruses have been carefully examined [6][7][8]. Specifically, the HIV-specific models [6] indicate an outstanding fit when applied to the HIV data in comparison to other general models. The influenza-specific model [8] was introduced by Dang et al., learned from millions of residues. Experimental results demonstrated a significant better in analyzing influenza protein sequences than other models. Similarly, rtREV [7] model was estimated for retroviruses.

Dengue virus (DENV), the cause of the life-threatening dengue hemorrhagic fever, re-emerged in the past decades, at a dangerous level, especially in tropical and subtropical regions [9]. According to a report from WHO [10], there are a total 390 million dengue infections per year and approximate 3.9 billion people in 128 countries are at risk of infection with dengue viruses. Because of the severity and emergency of the virus, intensive studies at the molecular level of Dengue viruses have been being deployed [11][12][13][14][15].

Our work focused on estimating an amino acid substitution model that best fits the evolution of dengue protein sequences and hence should be used for studies of dengue virus proteins. The rest of the paper is organized as follows: Theoretical background of amino acid substitution models is represented in the section II (Method). Section III (Results) describes the experiment results and the comparison among models. Conclusions and perspectives are given in the last section.

II. METHODS

Generally, amino acid sites are assumed to be independently substituted and the rate of substitution is remaining stable over time. We normally use a homogeneous, continuous, and stationary Markov process to model the substitution process between amino acids [16][17]. The model is characterised by an instantaneous substitution rate 20x20 matrix, denoted by $Q = \{q_{xy}\}$ where

XXX-X-5386-6113-0/18/$31.00 ©2018 IEEE

$q_{xy\ (x\neq y)}$ is the number of amino acid x change into y per one unit of time and q_{xx} is assigned to satisfy the stationary condition, i.e., $\sum q_{xy} = 0$ for any x. The frequencies of amino acids are also assumed to be stationary and represented by an equilibrium 20-element vector $\pi = (\pi_x)$, where π_x is the frequency of amino acid x.

As usual, we also assume that the substitution process is time-reversed, thus we can formulate Qas follows: $q_{xy} = \pi_y r_{xy}$ and $q_{xx} = -\sum_{x\neq y} q_{xy}$ where $r_{xy}$is the exchangeability coefficient between amino acids x and y. The coefficient matrix is symmetric, that is $r_{xy} = r_{yx}$.

The frequency vector π has 19 free parameters and can be directly approximated from the data, however, the rate matrix Q has 190 free parameters and much more difficult to be estimated from the data. In this study, we applied the maximum likelihood method to estimate Q. The estimation process consists of four main steps: Data pre-processing, Tree reconstruction, Model estimation, and Model comparison (see Fig. 1).

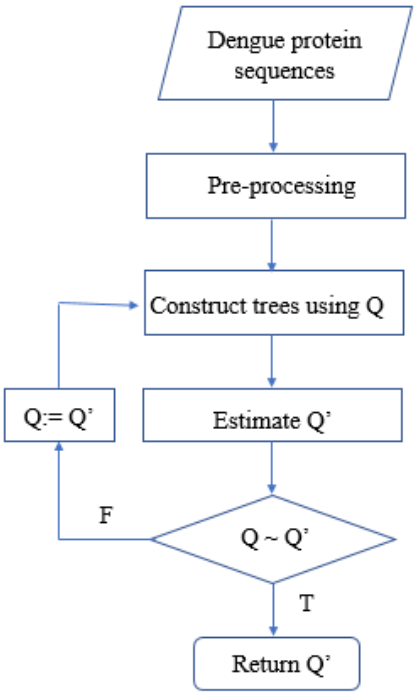

Fig. 1. The maximum likelihood-based process to estimate an amino acid substitution model for protein sequences of dengue viruses.

Step 1: Data pre-processing

The dengue viruses are members of the Flaviviridae family [11]. They have been first isolated in 1943 and stored at the NCBI (National Center for Biotechnology Information) since 1987. There are four different dengue virus types named DEN-1, DEN-2, DEN-3, and DEN-4. These four types are similar (share about 65% common genomes) and present all over the tropical and subtropical regions [11]. Their genomes contain a positive-sense RNA of about 11 kbs. This RNA encodes 3 structural proteins (C, M, E) and 8 non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, 2K, NS4B, NS5). To estimate an amino acid substitution model for dengue viruses, we downloaded all available amino acid sequences from https://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Data base/nph-select.cgi?taxid=12637. There are more than 20000 sequences in the database but many of these are identical. After removing the duplicated sequences, we obtained 10958 distinct sequences in which type 1 and type 2 take the major parts (39% and 32% respectively). To estimate a substitution model, the sequences of each virus type are divided into training and testing data sets containing 90% and 10% number of sequences, respectively. Table 1 describes the summary of data.

TABLE I. THE NUMBER OF DENGUE VIRUS AMINO SEQUENCES FOR FOUR TYPES AND 11 PROTEINS.

	DEN-1	DEN-2	DEN-3	DEN-4
C	1722	1432	936	356
M	1867	1614	1035	410
E	1739	1436	917	375
NS1	1678	1343	885	343
NS2A	1691	1345	883	338
NS2B	1671	1344	882	339
NS3	1669	1338	883	334
NS4A	1672	1341	887	328
NS2K	1673	1339	886	329
NS4B	1669	1316	885	327
NS5	1640	1298	867	325

Amino acid sequences were aligned using MUSCLE program [17]. Note that large alignments, containing of thousands of sequences, were divided into sub-alignments of at most 16 sequences using the tree-based splitting algorithm proposed in [18]. The splitting algorithm allows to estimate amino acid substitution models from large datasets.

Step 2: Tree reconstruction

We followed the maximum likelihood method described in [18] to estimate phylogenetic trees from multiple sequence alignments in the training dataset. Specifically, we used IQ-TREE program [19] to construct phylogenetic trees using the model Q. Note that LG model was assigned as the initial Q model.

Step 3: Model estimation

We employed an expectation–maximization algorithm, XRATE [20], to estimate a new model Q' using protein alignments and phylogenetic trees obtained from the previous steps.

Step 4: Model comparison

We compare the current model Q and newly estimated model Q'. If the difference between the new model Q' and the current model Q is not significant, then Q' is considered as the final best model estimated. Otherwise, Q is assigned by Q' and go to step 2. Experiment results show that the algorithm usually stops after three iterations.

III. RESULTS

We named the new amino acid substitution model for dengue viruses as DEN. We evaluated the fit of DEN on the training set and assessed the performance of DEN on the testing dataset in comparison to five other current models. The five models include four models for other viruses (FLU of influenza viruses, HIVb and HIVw of HIV viruses and rtREV of retro viruses) and LG (the current best general model [1]). We measured the difference of the log-likelihood and tree topologies constructed with DEN (M_1) and each of five models (M_2). We also used Pearson correlation to evaluate the correlations between the exchangeability matrices (frequency vectors) of M_1 and M_2 models.

A. *The fit of DEN model on training set*

Table 2 represents the likelihood improvements during the model training process. The values show a sharp rise of 54685 log likelihood unit after the first iteration. At the third iteration, the log likelihood only slightly increases, and the two matrices have a nearly 1 correlation, hence we stop the training at this point.

Look at the AIC (Akaike information criterion) measurement - an estimator of the relative quality of statistical models for a given set of data [21] it is obvious that AIC improvement of the final model (DEN) over the initial model (LG) is significant (i.e., 131162). It guarantees the worth of the DEN model over the penalty of 208 free parameters[21].

TABLE II. LOG-LIKELIHOOD OF THE TARGET FUNCTION ON THE TRAINING DATASET.

LG (initial model)	-6157890
First iteration	-6106205
Second iteration	-6093870
Third iteration (final model)	-6092301
AIC improvement	131162

B. *Likelihood improvement on testing dataset*

We evaluated the performance of DEN and other five models FLU, HIVb, HIVw, rtREV and LG by comparing the log-likelihood of trees which were inferred from testing alignments but with different models. To this end, IQ-TREE [19] was used to infer trees from 86 testing alignments. Fig. 2 shows the average log-likelihood difference per alignment between trees inferred with FLU, DEN, HIVb, HIVw, rtREV and those inferred with LG. Obviously, DEN is the best model while HIVw is the worst one. HIVb is the second best model (50 log-likelihood points lower than DEN model per alignment).

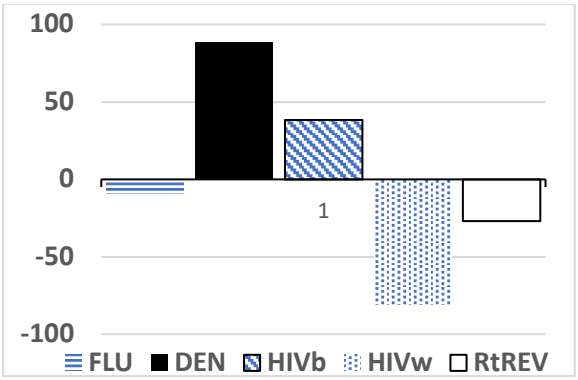

Fig. 2. Log-likelihood difference per alignment between LG model and other models. DEN is the best model on the testing dataset.

We also employed the Kishino-Hasegawa (KH) test [22] to check the statistical significance of the log-likelihood difference between trees constructed with two different models (denoted by M_1 and M_2). Table 3 represents the test results: The third column shows the number of alignments that M_1 is better (i.e., higher log-likelihood) than M_2; the fourth column shows the number of alignments that M_1 is significantly better than M_2 under KH test; the last column shows the number of alignments that M_2 is significantly better than M_1. We observed that DEN is significantly better that other models in most of the test alignments. For example, DEN is significantly better than LG and rtREV for all 86 alignments.

TABLE III. KISHINO-HASEGAWA (KH) TEST RESULT OF DEN AND OTHER MODELS WITH P-VALUE < 0.05

M_1	M_2	$\#M_1 > M_2$	$\#M_1 > M_2$ ($p < 0.05$)	$\#M_2 > M_1$ ($p < 0.05$)
DEN	LG	86	86	0
DEN	FLU	86	85	0
DEN	HIVb	86	83	0
DEN	HIVw	86	85	0
DEN	RtREV	86	86	0

C. *Model analysis*

We measured the correlations between DEN and other models. Table 4 shows that Dengue has an evolutionary pattern that considerably differs from other viruses such as HIV or Influenza due to the low correlation between their models (the highest correlation is only 0.884% which is between DEN and rtREV models).

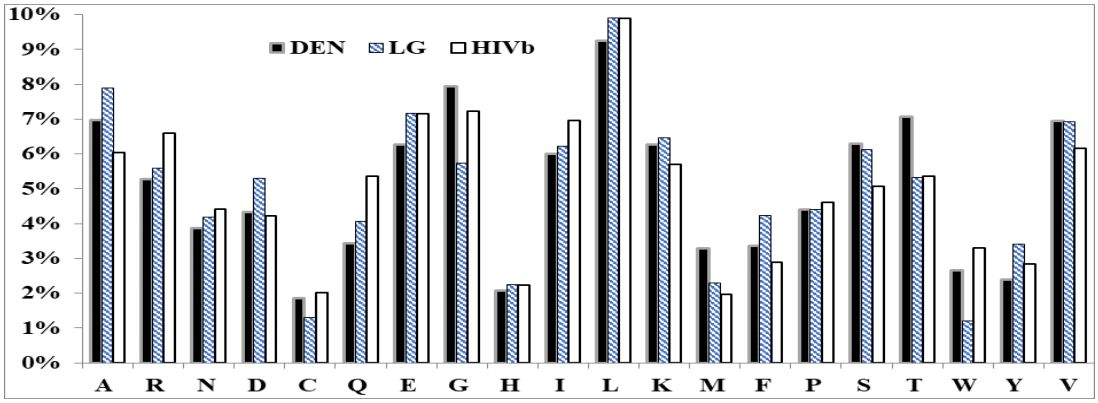

Fig. 3. *Amino acid frequencies of DEN, LG and HIVb models*

TABLE IV. THE PEARSON CORRELATIONS BETWEEN DEN AND OTHER MODELS

Model	Exchangeability matrix	Frequency vector
LG	0.954	0.908
RtREV	0.884	0.853
HIVb	0.878	0.896
FLU	0.846	0.803
HIVw	0.775	0.642

Figure 3 shows the difference between amino acid frequencies of DEN, LG and HIVb models. We observe some notable differences between frequencies of these models. For instance, the frequency of W (Tryptophan) in DEN (3%) is three times higher than that in LG (1%), while Q (Glutamine) frequency is only 3% in DEN, just over a half of that in HIVb.

The exchangeability coefficients of DEN, LG and HIVb models were plotted in Figure 4. In general, most values distributed in a similar trend due to biological constraints. For instance, isoleucine is frequently substituted by valine, methionine, leucine, threonine and phenylalanine, while other amino substitutions rarely happen as their corresponding coefficients are relatively small. Nevertheless, there are some remarkable differences. For example, the coefficients on T (Threonine) row are notably different among models. More specifically, the rate of amino acid T (Threonine) to amino acid A (Alanine) in HIVb model is about four times higher than that in DEN and LG models. Overall, exchangeability matrix and frequency vector of DEN are different from those of existing models.

D. The robustness of DEN model

The DEN model was estimated from the training dataset containing 90% of the dengue protein sequences. To examine the robustness of the DEN model, we estimated additional models from three other training datasets. Specifically,

- DENG: The model estimated from the training dataset consisting of all dengue protein sequences.
- DEN1: The model estimated from the training dataset containing the first half of all dengue protein sequences.
- DEN2: The model estimated from the training dataset containing the second half of all dengue protein sequences.

Table 5 shows extremely high correlations between exchangeability matrices of the four models. The smallest correlation value is 0.992 between DEN1 and DEN whilst DEN, DENG and DEN2 correlations are close to 1. Thus, the dengue dataset is sufficient enough to estimate a stable model.

TABLE V. CORRELATION BETWEEN EXCHANGEABILITY MATRICES OF DEN, DENG, DEN1, AND DEN2 MODELS

	DEN	DENG	DEN1	DEN2
DEN		1.000	0.992	0.999
DENG	1.000		0.992	0.998
DEN1	0.992	0.992		0.996
DEN2	0.999	0.998	0.996	

IV. CONCLUSIONS

Dengue virus infections are dangerous over the world with huge effect to public health. Despite thousands intensive studies of the virus at the genetic level have been conducted, challenging questions still remains. In this study, we proposed a new amino acid substitution model for dengue viruses. Experiments showed that DEN model is considerably different from existing models, and

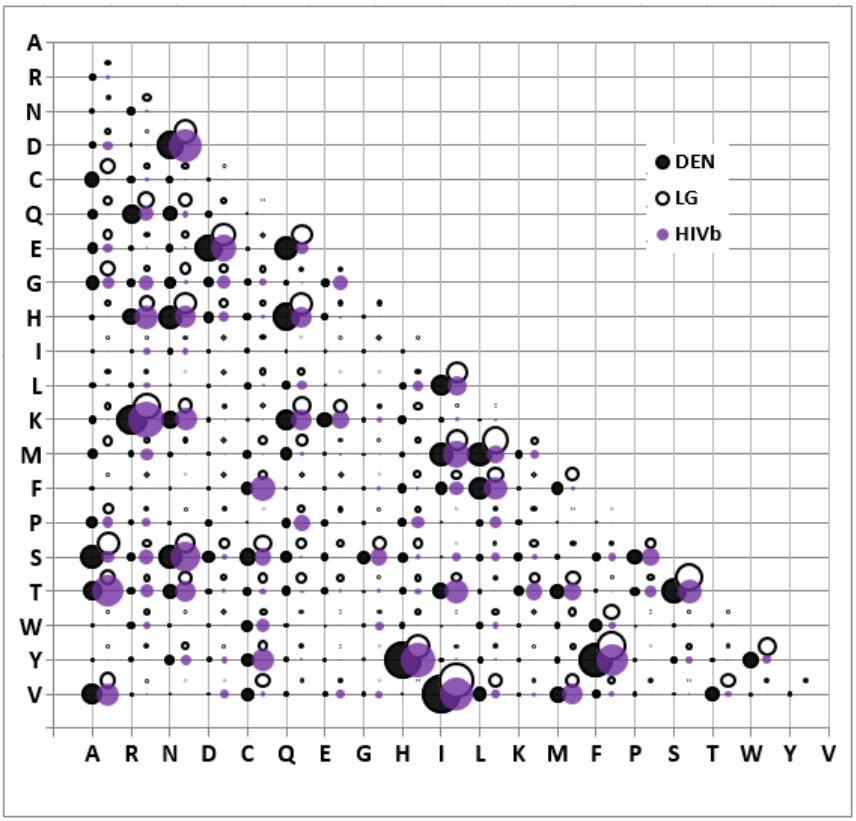

Fig. 4. The exchangeability coefficients in DEN, LG and HIVb models.

significantly outperforms other models when analyzing dengue protein sequences. We encourage researchers to use this new model for analyzing protein sequences of not only dengue viruses, but also other viruses in the Flaviviridae family.

REFERENCES

[1] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. (1981). 17:368–76.

[2] Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. (2008). 25:1307–20

[3] Dayhoff MO, Schwartz RM, Orcutt BC: A Model of Evolutionary Change in Proteins. Atlas of Protein Sequence Structure Volume 5. (1978)345-352.

[4] Jones DT, Taylor WR, Thornton JM: The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci (1992).275-282

[5] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol (2010). 59:307–321.

[6] Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, KosakovskyPond SL.HIV-Specific Probabilistic Models of Protein Evolution. PloS One (2007). 2:e503.

[7] Dimmic Ma, Rest J, Mindel D, Goldstein R. rtREV: An Amino Acid Substitution Matrix for Inference of Retrovirus and Reverse Transcriptase Phylogeny. J Mol Evol. (2002). 55. 65-73.

[8] Dang Cuong, Le Quang, Gascuel Olivier, Vinh Le. FLU, an amino acid substitution model for influenza proteins. BMC evolutionary biology. (2010). 10. 99. 10.1186/1471-2148-10-99.

[9] Igarashi, Akira. "Impact of dengue virus infection and its control." FEMS Immunology & Medical Microbiology 18.4 (1997): 291-300.

[10] http://www.searo.who.int/entity/vector_borne_tropical_diseases/data/data_factsheet/en/

[11] Beasley, D. W. C. & Barrett, A. D. T. Dengue: Tropical Medicine: Science and Practice, The Infectious Agent vol. 5, eds. G. Pasvol & S. L. Hoffman London: Imperial College Press. (2008). 29–74.

[12] Dowd KA, DeMaso CR, Pierson TC. Genotypic differences in dengue virus neutralization are explained by a single amino acid mutation that modulates virus breathing. MBio 6. (2015). e01559-15

[13] Guirakhoo, F. et al. A Single Amino Acid Substitution in the Envelope Protein of Chimeric Yellow Fever-Dengue 1 Vaccine Virus Reduces Neurovirulence for Suckling Mice and Viremia/Viscerotropism for Monkeys. Journal of Virology 78.18 (2004): 9998–10008.

[14] Drumond, Betania Paiva et al. Phylogenetic Analysis of Dengue Virus 1 Isolated from South Minas Gerais, Brazil. Brazilian Journal of Microbiology 47.1 (2016). 251–258.

[15] Shannon N. B. et al. Selection-Driven Evolution of Emergent Dengue Virus, Molecular Biology and Evolution, Vol. 20. (2003). 1650–1658.

[16] Le SQ, Dang CC, Gascuel O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol. (2012). 29: 2921–36

[17] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. (2004). 49:652-670

[18] Dang Cuong, Vinh Le, Gascuel Olivier, Hazes Bart, Le Quang. FastMG: a simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets. BMC bioinformatics (2014). 15. 341. 10.1186/1471-2105-15-341

[19] Nguyen LT, Schmidt H, von Haeseler A, Minh B. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. (2014). 32. 10.1093/molbev/msu300

[20] Peter S Klosterman, Andrew V Uzilov, Yuri R Bendaña, Robert K Bradley, Sharon Chao, Carolin Kosiol, Nick Goldman and Ian Holmes. XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics, vol. 7. (2006) p. 428.

[21] Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. (1974).19:716–23

[22] Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in homiboidea. J Mol Evol. (1989). 29: 170-179