VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name: Le Kim Thu 2. Sex: female
3. Date of birth: 17/10/1985 4. Place of birth: Hanoi
5. Decision of recognition of PhD students No. 841/QĐ-CTSV, dated 4/9/2018 of the Rector of the University of Engineering and Technology, Vietnam National University, Hanoi.
6. Changes in academic process
   - Extension of the education time according to the Decision No. 804/QĐ-ĐT and921/QĐ-ĐT of the Rector of the University of Engineering and Technology, Vietnam National University, Hanoi.
   - The thesis title is changed according to the suggestion of The Thesis Overall Evaluation Seminar Committee of the University of Engineering and Technology, 19/6/2023.

   Old thesis title: Optimizing amino acid substitution models for genomic data

   New thesis title: Developing amino acid substitution models for genomic data
7. Thesis title: Developing amino acid substitution models for genomic data
8. Major: Computer science 9. Code: 9480101.01
10. Supervisor: Assoc.Prof.Dr. Le Sy Vinh
11. **Summary of the new findings of the thesis**

    The thesis proposes a new amino acidsubstitution model, two partition alignment algorithms and a new method to compute specific evolutionary site rates. All proposals aimed at increase the accuracy of the reconstructed phylogeny for a given alignment.Specifically:

    (1) The thesis developed FLAVI – an amino acid substitution model specific to Flavivirus viruses. Experiments show that phylogenies reconstructed using FLAVI have higher likelihood values than existing models for Flavivirus data.

(2) The thesis proposes two algorithms mPartition and gPartition to partition large to genome size alignments. Both algorithmsuse substitution model and evolutionary rate at each site in partitioning process.

a. mPartition algorithm can be used for nucleotides and amino acids alignments. Experiments suggested that the maximum likelihood phylogeny trees constructed usingmPartition'spartitioning schemes had better AIC/BIC scores than the trees using other partitioning schemes in most of tested cases.

b. The gPartition algorithm can be used for nucleotide alignments. Test alignments which contain upto million sites was partitioned in less than 24 hours, meanwhile mPartition algorithm couldnot finish after 72 hours. Using the partitioning schemes generated by gPartitionalso reconstructed better maximum likelihood tree than without partitioning and than using the schemes of rate-based partitioning methods.

(3) In addition, the thesis proposesfastTIGER– a rapid method for estimating evolutionary rates of sites.fastTIGER has linear complexity according to the number of sites in alignment, therefore, it is more suitable for calculating the evolutionary rate for genome size aligments than the TIGER rate estimation algorithm.

12. Practical applicability, if any

The results in this thesis contribute to:

FLAVI model can be used to search protein sequences in databases and to reconstruct phylogeny trees. In the process of reconstructing a phylogeny, a partitioning scheme model (created by aligned partitioning algorithms) can be used to improves the accuracy of inffered tree.

fastTIGER is not only used in partitioning algorithms but also is used forpreprocessing data.

13. Further research directions, if any

- Find new method for partitioning a given alignment.
- Consider different constraints need to be satisfedof input data.

14. Thesis-related publications

- Le Kim Thu, Cuong Dang Cao, and Vinh Le Sy. 2018. "Building a specific amino acid substitution model for dengue viruses." 2018 10th International Conference on Knowledge and Systems Engineering (KSE) 242-246.
- Le Kim Thu, and Vinh Le Sy. 2020. "A protein alignment partitioning method for protein phylogenetic inference." 2020 RIVF International Conference on Computing and Communication Technologies (RIVF) 1-5.
- Le Kim Thu, Vinh Le Sy, Dong Do Duc, Thang Bui Ngoc, and Phuong Thao Nguyen Thi. 2020. "iK-means: an improvement of the iterative k-means partitioning algorithm." 2020 12th International Conference on Knowledge and Systems Engineering (KSE) 300-305.
- Le Kim Thu, and Vinh Le Sy. 2020. "FLAVI: An Amino Acid Substitution Model for Flaviviruses." Journal of molecular evolution 88 (5): 445-452.
- Le Kim Thu, and Vinh Le Sy. 2020. "mPartition: A Model-based method for partitioning alignments." Journal of Molecular Evolution 88 (8): 641-652.
- Le Kim Thu, and Vinh Le Sy. 2021. "fastTIGER: A rapid method for estimating evolutionary rates of sites from large datasets." 2021 13th International Conference on Knowledge and Systems Engineering (KSE).
- Le Kim Thu, and Vinh Le Sy. 2022. "A protein secondary structure-based algorithm for partitioning large protein alignments." 2022 14th International Conference on Knowledge and Systems Engineering (KSE) 1-5.
- Le Kim Thu, Diep Hoang Thi, Dong Do Duc, Thang Bui Ngoc, Phuong Thao Nguyen Thi, and Vinh Le Sy. 2022. "gPartition: An Efficient Alignment Partitioning Program for Genome Datasets." VNU Journal of Science: Computer Science and Communication Engineering.

Date:	*Date:*
Scientific advisor	PhD. Candidate
Signature:..	Signature: ....
Full name:...	Full name:..