Journal of Computer Science and Cybernetics, V.37, N.4 (2021), 351–363
DOI 10.15625/1813-9663/37/4/15937

MODELING AMINO ACID SUBSTITUTIONS FOR WHOLE GENOMES

LE SY VINH

University of Engineering and Technology, Vietnam National University, Hanoi

Abstract. Modeling amino acid substitution process is a core task in bioinformatics. New advanced sequencing technologies have generated huge datasets including whole genomes from various species. Estimating amino acid substitution models from whole genome datasets provides us unprecedented opportunities to accurately investigate relationships among species. In this paper, we review state-of-the-art computational methods to estimate amino acid substitution models from large datasets. We also describe a comprehensive pipeline to practically estimate amino acid models from whole genome datasets. Finally, we apply amino acid substitution models to build phylogenomic trees from bird and plant genome datasets. We compare our newly reconstructed phylogenomic trees and published ones and discuss new findings.

Keywords. Evolutionary analysis; Amino acid substitution models; Protein analysis; Genome analysis; Maximum likelihood estimation.

1. INTRODUCTION

The information of evolutionary relationships among species plays an important role in genomic studies such as determining gene functions, investigating genetic-related diseases, developing new drugs/vaccines, or understanding population history. Therefore, developing computational methods to analyze evolutionary relationships among sequences is a key problem in bioinformatics. To this end, the substitution process among nucleotides/amino acids during the evolution must be properly modeled. The nucleotide/amino acid substitution models are required in maximum-likelihood (ML) phylogenetic tree reconstructions or Bayesian methods to calculate the likelihood of data [1]. Note that the ML and Bayesian methods have been proved to be more accurate than other methods (e.g., maximum parsimony methods, distance-based methods) in analyzing the evolutionary relationships among species based on their genetic data [1]. Using nucleotide/amino acid substitution models helps the ML and Bayesian methods avoid systematic errors (i.e., back mutations, parallel mutations, multiple mutations) when constructing phylogenies from diverse datasets. The distance-based approach requires nucleotide/amino acid substitution models to estimate

Dedicated to Professor Phan Dinh Dieu on the occasion of his 85th birth anniversary.

*Corresponding author.
E-mail addresses: vinhls@vnu.edu.vn.

pairwise distances between sequences. Employing wrong nucleotide/amino acid substitution models results in incorrect phylogenetic trees. The amino acid substitution models such as PAM [2] or BLOSUM [3] can also play as score matrices in the sequence similarity search. The roles and applications of amino acid substitution models are summarized by [4].

A nucleotide substitution model comprises a 4×4 matrix representing instantaneous substitution rates among 4 nucleotides, i.e., Adenine (A), Thymine (T), Guanine (G), and Cytosine (C); and a nucleotide frequency vector describing nucleotide frequencies. A general time-reversible nucleotide substitution model has only 8 free parameters that can be directly estimated from nucleotide sequences under the study [1]. In short, estimating nucleotide substitution models is not a computationally expensive task.

Amino acid sequences are more conserved than nucleotide sequences, therefore, they are frequently used to construct evolutionary relationships among species, especially among distantly related-species. Modeling amino acid substitutions is orders of magnitude more challenging than modeling nucleotide substitutions. An amino acid substitution model comprises a 20× 20 instantaneous substitution rate matrix and an amino acid frequency vector of the 20 amino acids. The time-reversible amino acid substitution model contains 208 parameters, therefore, small or medium size datasets normally do not provide enough phylogenetic signals for correctly estimating such large number of parameters. To solve this problem, amino acid substitution models are empirically estimated from large protein datasets in advance.

Computational methods have been proposed to estimate amino acid substitution models since 1970s. The counting methods calculate the number of observed changes between amino acids in a set of protein sequences to estimate substitution rates among amino acids. They were applied to small protein datasets to estimate Dayhoff (PAM) models [2], or latter JTT model [5] or BLOSUM model [3]. Note that the counting methods are only applicable for closely-related protein datasets.

Maximum likelihood methods have been proposed to estimate amino acid substitution models from diverse datasets. Whelan and Goldman applied the maximum likelihood method to estimate WAG model from a dataset of 3,905 globular amino acid sequences [6]. Experiments showed that the WAG model outperformed the Dayhoff and JTT models in building maximum-likelihood trees. Approximate maximum likelihood approach has been developed to estimate amino acid substitution models from larger datasets [7-10]. Note that amino acid substitution models that were estimated from general protein datasets such as WAG [6], LG [7] or Q.pfam [9] are called general amino acid substitution models.

Among available amino acid substitution models, PAM and BLOSUM are frequently used as score matrices in the sequence similarity search. However, as the models were constructed from a limited number of closely-related protein sequences, they are normally not as good as newly estimated general models such as Q.pfam in analyzing large and diverse protein datasets.

The general amino acid substitution models are suitable for analyzing general protein sequences, however, they do not properly describe amino acid substitutions of viruses due their rapid evolution. A number of models have been estimated for different viruses such as HIV models for HIV viruses [11]; FLU model for influenza viruses [12] or FLAVI model for flaviviruses [13]. Experiments revealed that the virus-specific models were much better than general models in analyzing amino acid data from viruses.

The next generation sequencing technologies have produced a huge amount of nucleotide

and amino acid sequences. Thousands of genomes from various species give us unprecedented opportunities to investigate the evolutionary relationships among species from the genome point of views. The current model estimation methods are not designed to work on whole genome datasets due to the computational expense of estimating a large number of parameters and the heterogeneity of evolutionary rates among genes. In other words, methods to analyze whole genome datasets must be able to handle thousands of genes with the variability of evolutionary rates among genes.

2. METHODS

2.1. Amino acid substitution model

The substitutions between amino acids during the evolution is typically modeled by a time-homogeneous, time-continuous, and stationary Markov process [1, 6, 7, 9]. An amino acid substitution model M comprises an instantaneous substitution rate matrix $Q = \{q_{x,y}\}$ representing instantaneous substitution rates between amino acids, and an amino acid frequency vector $\boldsymbol{\pi} = \{\pi_x\}$. Specifically, the model has four conditions:

- The substitution rate from amino acid x to amino acid y is independent of the substitution history of amino acid x (Markov property).
- The substitution rates among amino acids are constant over time (time-homogeneous).
- The substitutions between amino acid can occur at any time in the process (time-continuous).
- The frequencies of amino acids are at equilibrium (stationary).

The matrix Q and vector $\boldsymbol{\pi}$ are dependent and $Q\boldsymbol{\pi} = 0$. Technically, Q can be decomposed into $q_{x,y} = \pi_y r_{x,y}$ and $q_{x,x} = -\sum_{x \neq y} q_{x,y}$ where $R = \{r_{x,y}\}$ is the exchangeability rate matrix between amino acids. The matrix R contains 380 parameters (20 entries on the main diagonal can be determined from off-diagonal entries). Typically, we assume that the amino acid substitution process is time-reversible, i.e., exchangeability rate from amino acid x to amino acid y is required to be the same as the exchangeability rate from y to x. Thus, $r_{x,y} = r_{yx}$ and Rconsists of 190 parameters.

Given an instantaneous substitution rate matrix Q, the transition probability matrix $P(t) = \{p_{xy}(t)\}$ in which $p_{xy}(t)$ is the probability to change from amino acid x to amino acid y during the evolutionary time t can be computed as following

$$p_{xy}(t) = e^{Qt}.$$

The probability $P(t)$ is used to calculate the likelihood of data in phylogenetic analysis. Note that the instantaneous substitution rate matrix Q is normally scaled such that the expected number of substitutions per time unit is one (i.e., $-\sum \pi_x q_{xx} = 1$). As a result, $p_{xy}(t)$ is the probability to change from amino acid x to amino acid y after t substitutions.

2.2. Model estimation

Till now, maximum likelihood is the best approach to estimate amino acid substitution models. A time-reversible amino acid substitution model M consists of 208 parameters

that must be estimated from a number of protein (amino acid) alignments to overcome the overfitting problem. Given a dataset $\boldsymbol{A} = (A_1, \ldots, A_n)$ of n protein alignments, the maximum-likelihood method estimate a model M to maximize the likelihood value $L(\boldsymbol{A}|M)$.

We assume that protein alignments are independent, thus, the likelihood value $L(\boldsymbol{A}|M)$ can be calculated from individual alignments as following

$$L(\boldsymbol{A}|M) = \prod_{i=1}^{n} L(A_i|M),$$

where $L(A_i|M)$ is the likelihood of alignment A_i. To calculate $L(A_i|M)$ we have to determine phylogenetic tree T_i for alignment A_i. Let $\boldsymbol{T} = (T_1, \ldots T_n)$ be the tree set corresponding to the dataset $\boldsymbol{A}$, i.e., T_i is the corresponding phylogenetic tree constructed from alignment A_i. The likelihood $L(\boldsymbol{A}|M)$ can be calculated as following

$$L(\boldsymbol{A}|M) = \prod_{i=1}^{n} L(A_i|M) = \prod_{i=1}^{n} L(A_i|M, T_i).$$

The heterogeneity of amino acid substitution rates among sites can be modeled by a site rate model V. Typically, a site rate model V combines a gamma distribution model Γ and an invariant rate model I [14]. We also assume that amino acid substitutions among sites are independent, thus, the likelihood value $L(\boldsymbol{A}|M, V, \boldsymbol{T})$ can be specifically calculated as following

$$L(\boldsymbol{A}|M, V, \boldsymbol{T}) = \prod_{i=1}^{n} L(A_i|M, V, T_i) = \prod_{i=1}^{n} \prod_{j=1}^{l_i} L(A_{ij}|M, V, T_i),$$

where l_i is the length of alignment A_i, and A_{ij} is the data at site j of alignment A_i. The likelihood value $L(A_{ij}|M, V, T_i)$ can be calculated by the conditional probability $P(A_{ij}|M, V, T_i)$ of data A_{ij} given substitution model M, site rate model V and tree T_i.

The maximum likelihood estimation method determines parameters of amino acid substitution model M (together with site rate model V and tree set $\boldsymbol{T}$) to maximize the likelihood value $L(\boldsymbol{A}|M, V, \boldsymbol{T})$.

2.3. Approximate maximum likelihood estimation

Finding the maximum likelihood value $L(\boldsymbol{A}|M, V, \boldsymbol{T})$ is computationally difficult because we have to simultaneously estimate a large number of parameters of substitution model M, site rate model V, and especially tree set $\boldsymbol{T}$. Approximate maximum likelihood methods have been proposed to estimate parameters of a model M from large datasets [6, 7, 10]. The methods revealed that the parameters of model M can be accurately estimated using nearly optimal trees $\boldsymbol{T}$ and site rate model V. Thus, we can iteratively estimate parameters of model M, trees set $\boldsymbol{T}$, and site rate model V instead of estimating them simultaneously.

Given a set of protein alignments $\boldsymbol{A}$, the approximate maximum likelihood method includes four main steps as follows (see Figure 1):

- **Initial step:** Initialize the current best-fit model Q by the most proper available model, e.g., LG for general dataset or FLU for virus dataset. The amino acid frequency vector π can be estimated by counting directly from protein alignments $\boldsymbol{A}$.

- **Building tree step:** Determine maximum likelihood trees $\boldsymbol{T}$ and site rate model V for alignments $\boldsymbol{A}$ based on the current best-fit model Q.
- **Estimating model step:** Estimating parameters of a new model Q' based on alignments **A,** maximum likelihood trees $\boldsymbol{T}$ and site rate model V.
- **Stopping step:** If Q and Q' are highly correlated, stop the estimation process and consider Q' as the best-fit model for the dataset. Otherwise, assign Q by Q' and go to the Building tree step.

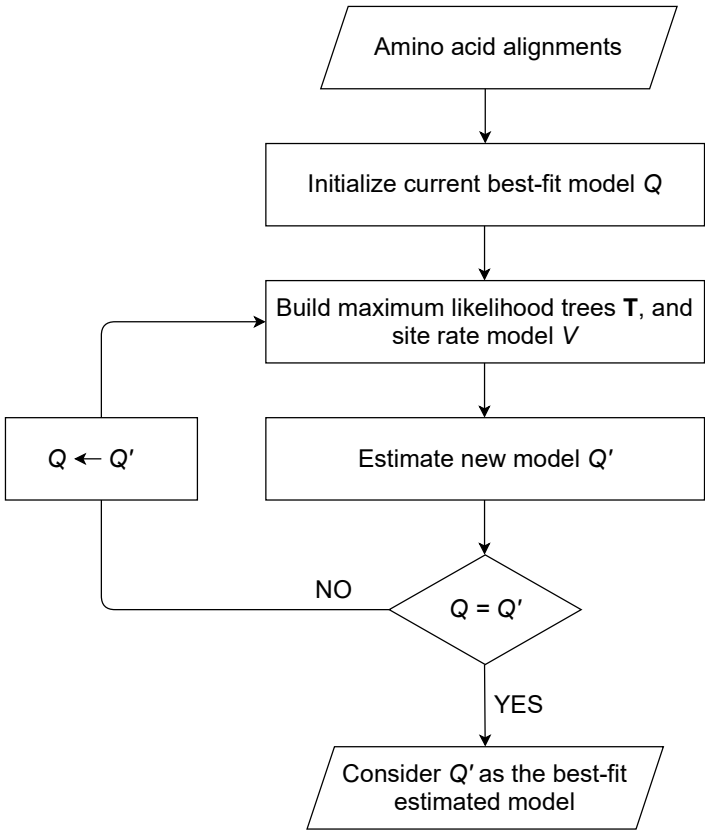

Figure 1: The approximate maximum likelihood method to estimate an amino acid substitution model from a set of amino acid alignments

2.4. Fast maximum likelihood estimation method

The approximate maximum likelihood method is still computational expensive when working on alignments with a large number of sequences due to the step of building maximum likelihood trees. For example, building maximum likelihood trees took 319 out of 324 hours to estimate an amino acid model from a dataset of 100 alignments in the HSSP database [8]. Although a number of phylogenetic reconstruction methods such as IQPNNI [15, 16], IQ-TREE [17], PhyML [18] or RaxML [19] have been proposed to efficiently build maximum likelihood trees from large alignments, they are not efficient enough to build thousands of maximum likelihood trees to estimate amino acid substitution models.

To overcome the computational burden in building maximum likelihood trees from large alignments, alignment splitting methods have been proposed to divide large alignments into smaller alignments such that the smaller alignments still contain enough phylogenetic signals to estimate amino acid substitution models. The maximum likelihood trees will be constructed from smaller alignments instead of large alignments. A fast and accurate procedure, called FastMG, has been proposed to combine the alignment splitting algorithm and the approximate maximum likelihood method to estimate amino acid substitution models from large datasets [8].

The FastMG procedure includes two main phases: alignment splitting and model estimating. The alignment splitting phase applies the tree-based alignment splitting algorithm to divide large alignments into smaller alignments such that each smaller alignment contains at most k sequences and at least $k/2$ sequences. The k value is determined to compromise the computational expense in building maximum likelihood trees and the amount of phylogenetic signals for correctly estimating substitution rates among amino acids. The model estimating phase employs the approximate maximum likelihood method to estimate amino acid substitution models from smaller alignments instead of large original alignments.

The FastMG procedure was examined on large datasets, i.e., HSSP dataset and Pfam dataset. Experiments showed that the FastMG procedure was an order of magnitude faster than the approximate maximum likelihood method, and more importantly there was no apparent loss in the quality of the estimated model. Note that $k = 16$ was recommended to compromise the estimation time and model quality.

2.5. Model assessment

Phylogenetic trees represent the evolutionary relationships among species where nodes represent species and branch lengths indicate the divergence in terms of the number of substitutions or the evolutionary time between two species. The phylogenetic trees are typically constructed based on either nucleotide or amino acid data. The amino acid data are more conserved than the nucleotide data, therefore more preferable to study the evolutionary relationships among diverse species.

A number of amino acid substitution models have been estimated from large empirical amino acid datasets for different data types. The model finder method should be applied to select the best fit model for the data under the study [20]. The Akaike information criterion (AIC) [21] and Bayesian information criterion (BIC) [22] can be used to measure the fit of a model to the data. The AIC and BIC scores are calculated from the likelihood value of the constructed tree and the penalty of free parameters in the models used to build the tree. The better AIC/BIC score indicates the better model in building the maximum likelihood tree.

3. ESTIMATING AMINO ACID SUBSTITUTION MODELS FROM WHOLE GENOMES

3.1. Genome partitioning

The first step in estimating amino acid substitution models from whole genomes is to divide genomes into separated partitions (loci or subsets) such that sites in the same partition

are assumed to evolve under the same amino acid substitution process (substitution models, site rate models). This can be done by using prior knowledges such at gene boundaries or codon positions to group sites into subsets [14, 23-25].

Dividing genomes into separated loci by prior knowledges has two limitations: 1) the gene boundary or codon information are not always available for genomes under the study; 2) it is not biologically plausible to assume that all sites of the same gene or at the same codon position are under the same evolutionary process. They might evolve at different rates and follow different substitution patterns. To handle the problem, computational methods to divide genomes into proper partitions have been proposed [25-29]. The alignment partitioning methods classify sites into several disjoint partitions (a partition scheme) based on their site rates and amino acid substitution patterns.

The site rate-based partitioning methods group sites with similar evolutionary rates into the same subset [25-27]. The methods work well on a number of datasets, however, they suffer a critical pitfall that they group all invariant sites (or nearly invariant sites) into one subset. The partition of all invariant sites significantly increases the likelihood of data but results in biased trees [28, 30].

Amino acid sites with similar evolutionary rates might follow different amino acid substitution patterns. We need to combine both site rates and amino substitution patterns to classify sites such that sites in the same partition have similar evolutionary rates and substitution patterns (amino acid substitution model). The mPartition method [29] has been recently proposed with three key ingredients: site rate-based partitioning; model-based partitioning; and invariant site partitioning. First, the mPartition algorithm classifies sites into different partitions based on their evolutionary rates, i.e., sites in the same partition have similar evolutionary rates. Second, the mPartition method determines the best-fit amino acid substitution model for each partition. The substitution models are used to re-classify sites into partitions based on their best-fit models, i.e., sites are re-classified into partitions with the highest likelihood values. Finally, the mPartition algorithm was designed to classify invariant sites into different partitions proportionally to their likelihood values with respect to the partitions to avoid the pitfall of site rate-based partitioning methods. Experiments on both real and simulated datasets showed that the mPartition method outperformed site rate-based methods in both constructing true trees and maximum likelihood trees.

3.2. Model heterogeneity

Using the same site rate model V for all alignments/partitions is not biologically realistic. To solve the problem, a set of site rate models $\boldsymbol{V} = (V_1, \ldots, V_n)$ for alignments $\boldsymbol{A}$, i.e., V_i is the site rate model for alignment A_i are determined during the model estimation procedure. Typically, the site rate model V_i combines a gamma distribution model and an invariant rate model [14]. The distribution-free rate models with several site rate categories can be also employed to better describe the rate heterogeneity among sites [20].

The approximate maximum likelihood approach using the same amino acid substitution model Q to build maximum likelihood trees for all alignments must be improved, i.e., each alignment/partition of the genome should be analyzed with its best-fit substitution model. To this end, we revise the Building tree step such that the best-fit substitution model for each alignment will be selected from a set of published models and the currently estimated model Q. Let $\boldsymbol{M} = \{M_1, M_2, \ldots, M_n\}$ be the set of substitution models where M_i is the best-fit

substitution model for alignment A_i. Note that M_i can be either the currently estimated model Q or one of published models. We build the maximum likelihood tree T_i for alignment A_i using site rate model V_i and substitution model M_i.

The QMaker algorithm was developed to include both site rate models $\boldsymbol{V}$ and substitution models $\boldsymbol{M}$ into the model estimation process [9]. It consists of five main steps:

- **Initial step:** Initialize the current best-fit model Q by the most proper published model (e.g., LG for general datasets or FLU for virus datasets).

- **Model finder step:** For each alignment A_i, select the best-fit substitution model M_i (chosen from published models and the current best-fit model Q), and the site rate model V_i.

- **Tree building step:** For each alignment A_i, construct tree T_i based on the site rate model V_i and substitution model M_i.

- **Model estimation step.** Estimate parameters of new model Q' from all alignments $\boldsymbol{A}$ using maximum likelihood trees $\boldsymbol{T}$, site rate models $\boldsymbol{V}$, and substitution models $\boldsymbol{M}$.

- **Stopping step:** If Q and Q' are highly correlated, stop the estimation process and consider Q' as the best-fit model for the dataset. Otherwise, assign Q by Q' and go to the Model finder step.

Experiments on whole genome datasets showed that QMaker was better than other model estimation methods. The amino acid substitution models estimated from whole genome datasets significantly outperformed available models in building maximum likelihood trees.

Figure 2 describes an overview scheme for estimating amino acid substitution models from genome datasets. Note that we should apply the alignment splitting algorithms to divide alignments with a large number of sequences into smaller alignments to avoid computational obstacle.

4. PHYLOGENOMIC TREES

We applied amino acid substitution models estimated from whole genome datasets [9] to build phylogenomic trees for plants and birds. The Plant dataset consists of 1308 loci with more than 430 thousands amino acid sites from 38 plant species [31]. The phylogenomic tree for plants was constructed using IQ-TREE 2 software [32] using newly estimated amino acid substitution model Q.plant [9]. Figure 3 shows that our phylogenomic tree and the published tree [31] have the same topology. Most of branches on both trees have high bootstrap support values indicating that the tree topologies are highly reliable.

The Bird dataset contains 8295 loci with more than 4.5 million amino acid sites from 48 bird species [33]. The phylogenomic tree for birds was constructed using the newly estimated model Q.bird [9]. Figure 4 shows that our phylogenomic tree has different topology in comparison with the published tree (i.e., red branches indicate the difference between two tree topologies). For example, the Pigeon is grouped together with Yellow-throated-sandgrouse in the published tree while it is grouped with Common cuckoo in our phylogenomic tree.

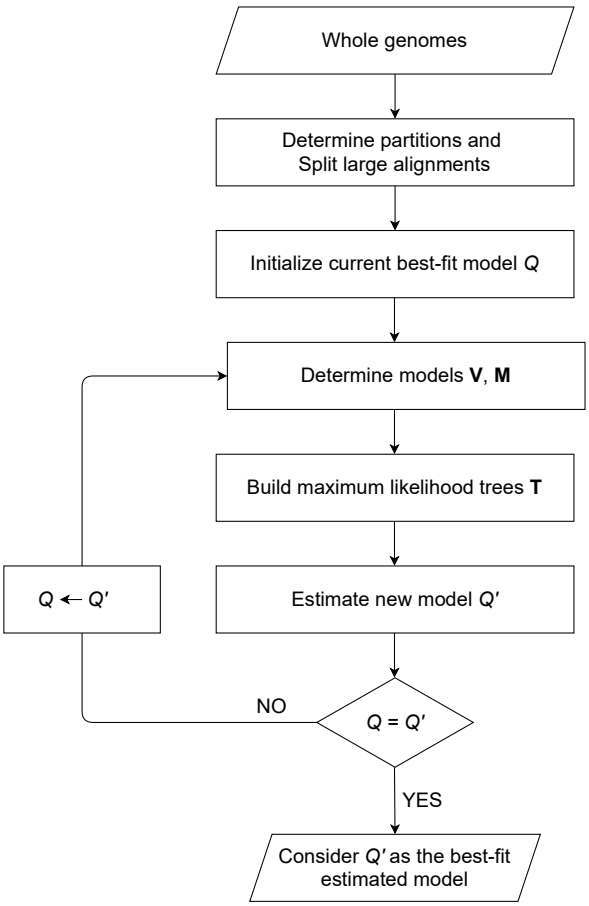

Figure 2: The overall scheme to estimate amino acid substitution models from whole genome datasets

5. CONCLUSIONS

The advanced sequencing technologies have created a huge amount of genomic data from different species. Studying the relationships among species from amino acid sequences requires amino acid substitution models. The time-reversible amino acid substitution models have a large number of parameters, therefore, they should be estimated from large/genome datasets.

We analyzed the state-of-the-art algorithms to estimate time-reversible amino acid substitution models from large datasets. As estimating amino acid substitution models is a complex process, each algorithm was designed to solve or improve some step in the estimation process. The approximate maximum likelihood algorithms can be efficiently used to estimate trees and models iteratively to reduce computational burden. The mPartition algorithm should be used to partition genomes into different partitions such that sites in each partition evolve under similar evolutionary models. The QMaker method is recommended to fully incorporate the heterogeneity of site rate and substitution models into the model

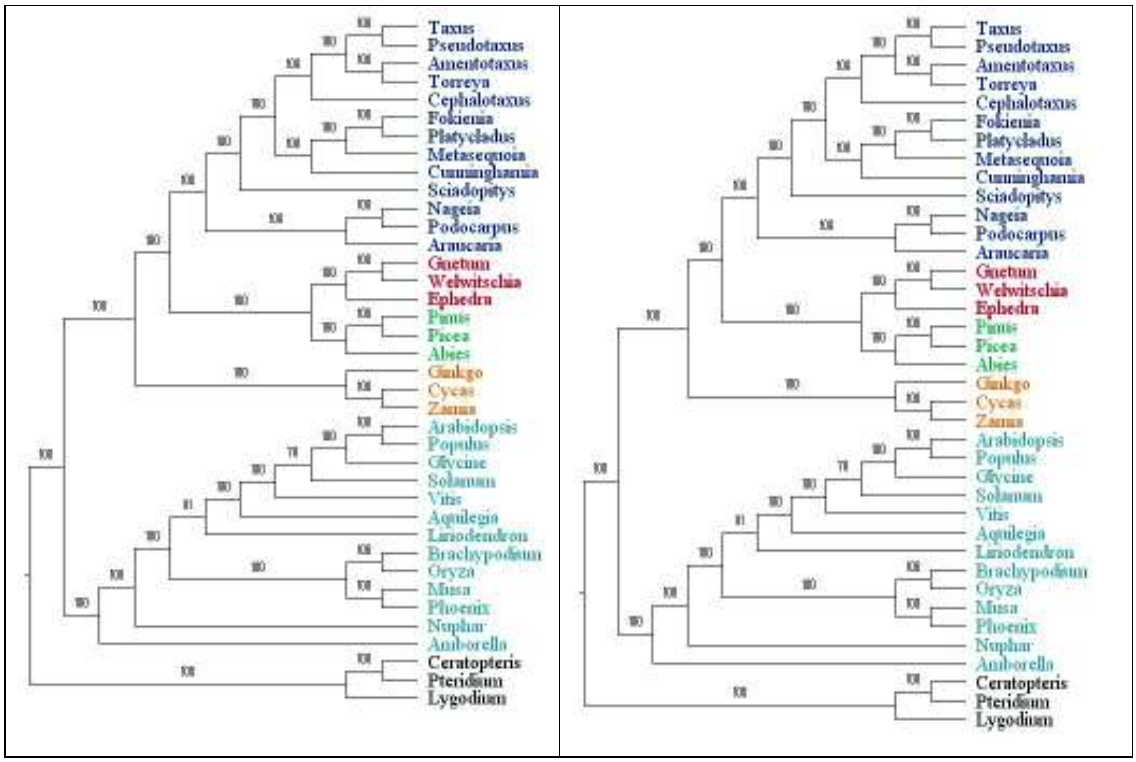

Figure 3: The published tree of 38 plant species (left) obtained from the paper [31]. Our phylogenomic tree constructed from the Plant genome dataset using the Q.plant model (right). The numbers on branches are bootstrap support values.

estimation process (i.e., each partition should be analyzed with its best-fit site rate and substitution models). Notably, the FastMG algorithm can be applied to split alignments with a large number of sequences into smaller alignments to significantly reduce the running time to build maximum likelihood phylogenetic trees. We strongly recommend researchers to follow our model estimation scheme in Figure 2 to efficiently estimate amino acid substitution models from genome datasets.

Finally, we constructed phylogenomic trees from Bird and Plant genome datasets including thousand loci with several million amino acid sites. The constructed phylogenomic tree for birds and the published one have identical topology with high bootstrap support values indicating their high reliability. However, the constructed phylogenomic tree for plants shows several topological differences in comparison with the published tree. The new and interesting findings require further assessments from biologists.

Up to date, amino acid substitution models are assumed to be time-reversible, i.e., backward and forward substitution rates are equally likely. The time-reversible assumption helps reduce the complexity in estimating amino acid substitution models, however, it is not biologically realistic. The assumption does not allow us to identify the root of tree, a central interest in phylogenetic studies. Two key computational obstacles in building time-nonreversible models are estimating a large number of parameters and constructing rooted maximum likelihood trees. These computational obstacles remain challenges for researchers in this field.

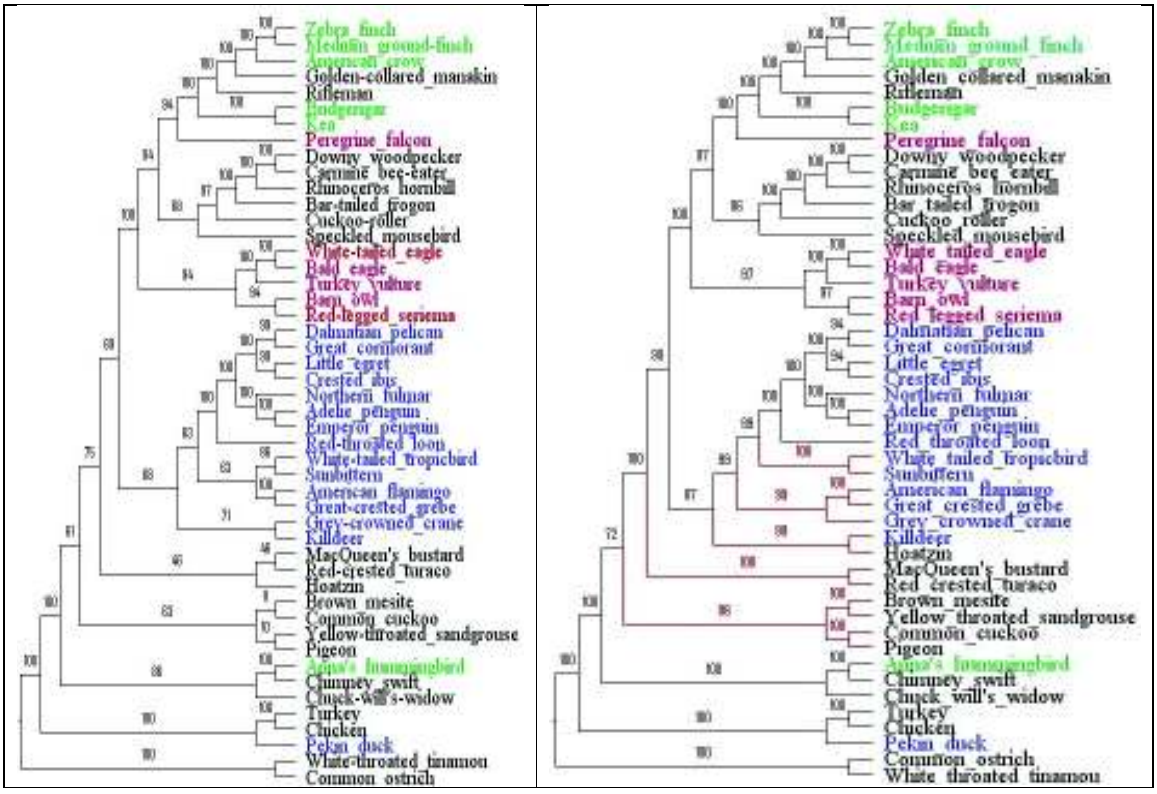

Figure 4: The published tree of 48 birds (left) obtained from the paper [33]. Our phylogenomic tree constructed from the Bird dataset using the Q.bird model (right). The numbers on branches are bootstrap support values.

6. ACKNOWLEDGEMENTS

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

REFERENCES

[1] Lemey Philippe, Marco Salemi, and Anne-Mieke Vandamme, "The Phylogenetic Handbook," *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. 2nd ed. Cambridge University Press. 2009 https://doi.org/10.1017/cbo9780511819049

[2] Dayhoff Mo, and RM Schwartz, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 22, pp. 345–52, 1978 https://doi.org/10.1.1.145.4315

[3] S. Henikoff, and J.G. Henikoff, "Amino acid substitution matrices from protein blocks," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915–10919, 1992. https://doi.org/10.1073/pnas.89.22.10915

[4] Thorne L. Jeffrey, "Models of protein sequence evolution and their applications," *Current Opinion in Genetics and Development*, vol. 10, pp. 602–605, 2000. https://doi.org/10.1016/S0959-437X(00)00142-8

[5] Jones T. David, R. William Taylor, and M. Janet Thornton, "The rapid generation of mutation data matrices from protein sequences," *Bioinformatics*, vol. 8, pp. 275–282, 1992 https://doi.org/10.1093/bioinformatics/8.3.275

[6] Simon Whelan, Nick Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach," *Molecular Biology and Evolution*, vol. 18, pp. 691–699, 2001. https://doi.org/10.1093/oxfordjournals.molbev.a003851

[7] Le Si Quang, and Olivier Gascuel, "An improved general amino acid replacement matrix," *Molecular Biology and Evolution*, vol. 25, no. 7, pp. 1307–1320, 2008. https://doi.org/10.1093/molbev/msn067.

[8] C.C. Dang, V.S. Le, O. Gascuel, et al., "FastMG: A simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets," *BMC Bioinformatics,* vol. 15, no. 341, 2014. https://doi.org/10.1186/1471-2105-15-341

[9] Bui Minh, Cuong Dang, Vinh Le, and Robert Lanfear, "QMaker: Estimating empirical models of protein evolution from large collections of alignments," *Systematic Biology*, 2021. https://doi.org/www.biorxiv.org/content/10.1101/2020.02.20.958819v1

[10] C.C. Dang, Vincent Lefort, Vinh Sy Le, Quang Si Le, and Olivier Gascuel, "Replacementmatrix: A web server for maximum-likelihood estimation of amino acid replacement rate matrices," *Bioinformatics*, vol. 27, no. 19, pp. 2758–2760, 2011. https://doi.org/10.1093/bioinformatics/btr435

[11] Nickle David C., Laura Heath, Mark A. Jensen, Peter B. Gilbert, James I. Mullins, and Sergei L. Kosakovsky Pond, "HIV-specific probabilistic models of protein evolution," *PLoS ONE*, vol. 2, no. 6, e503, 2007. https://doi.org/10.1371/journal.pone.0000503

[12] C.C. Dang, Q.S. Le, O. Gascuel, et al., "FLU, an amino acid substitution model for influenza proteins," *BMC Evolutionary Biology*, vol. 10, no. 99, 2010. https://doi.org/10.1186/1471-2148-10-99

[13] Le Thu Kim, and Le Sy Vinh, "FLAVI: An amino acid substitution model for flaviviruses," *Journal of Molecular Evolution*, vol. 88, pp. 445–452, 2020. https://doi.org/10.1007/s00239-020-09943-3

[14] Z. Yang, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, " *Molecular Biology and Evolution*, vol. 10, no. 6, pp. 1396–1401, 1993. https://doi.org/10.1093/oxfordjournals.molbev.a040082

[15] Vinh Le Sy, and Arndt Von Haeseler, "IQPNNI: Moving fast through tree space and stopping in time," *Molecular Biology and Evolution*, vol. 21, no. 8, pp. 1565–1571, 2004. https://doi.org/10.1093/molbev/msh176

[16] Minh, Bui Quang, Le Sy Vinh, Arndt von Haeseler, and Heiko A. Schmidt, "pIQPNNI: Parallel reconstruction of large maximum likelihood phylogenies," *Bioinformatics*, vol. 21, no. 19, pp, 3794–3796, 2005. https://doi.org/10.1093/bioinformatics/bti594

[17] Nguyen Lam Tung, Heiko A. Schmidt, Arndt Von Haeseler, and Bui Quang Minh, "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Molecular Biology and Evolution*, vol. 32, no. 1, pp. 268–274, 2015. https://doi.org/10.1093/molbev/msu300

[18] Guindon, Stéphane, and Olivier Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003. https://doi.org/10.1080/10635150390235520

[19] Stamatakis, Alexandros, "Using RAxML to Infer Phylogenies," *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 51: 6.14.1-6.14.14. https://doi.org/10.1002/0471250953.bi0614s51

[20] S. Kalyaanamoorthy, B. Minh, T. Wong, et al., "ModelFinder: Fast model selection for accurate phylogenetic estimates," *Nat Methods,* vol. 14, pp. 587–589, 2017. https://doi.org/10.1038/nmeth.4285

[21] H. Akaike, "A new look at the statistical model identification," in *IEEE Transactions on Automatic Control,* vol. 19, no. 6, pp. 716–723, December 1974. Doi: 10.1109/TAC.1974.1100705

[22] Schwarz, Gideon. 1978. "Estimating the dimension of a model," *The Annals of Statistics,* vol. 6, no. 2, pp. 461–464, 1978.

[23] M.C. Brandley, A. Schmitz, T.W. Reeder, "Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of Scincid Lizards," *Syst. Biol.*, vol. 54, pp. 373–90, 2005.

[24] N. Lartillot, H. Philippe, "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process," *Molecular Biology and Evolution,* vol. 21, no. 6, pp. 1095–1109, 2004. https://doi.org/10.1093/molbev/msh112

[25] J. Nylander, F. Ronquist, J. Huelsenbeck, and J. Nieves-Aldrey, "Bayesian phylogenetic analysis of combined data," *Systematic Biology*, vol. 53, no. 1, pp. 47–67, 2004. https://doi.org/10.1080/10635150490264699

[26] R. Lanfear, B. Calcott, S Ho, and S Guindon, "PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses," *Molecular Biology and Evolution*, vol. 29, no. 6, pp. 1695–1701, 2012. https://doi.org/10.1093/molbev/mss020

[27] P.B. Frandsen, B. Calcott, C. Mayer, et al., "Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates," *BMC Evol Biol,* vol. 15, no. 13, 2015. https://doi.org/10.1186/s12862-015-0283-7

[28] J. Rota, T. Malm, N. Chazot, C. Peña, N. Wahlberg, "A simple method for data partitioning based on relative evolutionary rates," *PeerJ,* vol. 6, e5498, 2018. https://doi.org/10.7717/peerj.5498

[29] T. Le Kim, V. Le Sy, "mPartition: A model-based method for partitioning alignments," *J Mol Evol,* vol. 88, pp. 641–652, 2020. https://doi.org/10.1007/s00239-020-09963-z

[30] S.M. Baca, E.F.A Toussaint, and K.B. Miller, "Molecular phylogeny of the aquatic beetle family noteridae (coleoptera: adephaga) with an emphasis on data partitioning strategies," *Molecular Phylogenetics and Evolution*, vol. 107, pp. 282–92, 2017.

[31] J-H. Ran, T-T. Shen, M-M. Wang, and X-Q. Wang, "Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between gnetales and angiosperms, " in *Proceedings of the Royal Society B* 285 (1881), 2018.

[32] Bui Quang Minh, Heiko Schmidt, Olga Chernomor, Dominik Schrempf, Michael Woodhams5, Arndt von Haeseler, and Robert Lanfear, "IQ-TREE 2: New models and efficient methods for phylogenetic inference in the Genomic Era," *Molecular Biology and Evolution*, vol. 37, no. 5, 2020.

[33] Jarvis D. Erich, Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y.W. Ho, et al., "Phylogenomic analyses data of the avian phylogenomics project," *GigaScience*, vol. 4, no. 1, December 2015, s13742-014-0038-1. https://doi.org/10.1186/s13742-014-0038-1

Received on March 15, 2021
Accepted on July 30, 2021