*Journal of Computer Science and Cybernetics, V.40, N.1 (2024), 53–66*
DOI no. 10.15625/1813-9663/19324

# ESTIMATING AMINO ACID SUBSTITUTION MODELS AND ROOTING BACTERIAL TREES
## SUBTITLE

CUONG CAO DANG, LE SY VINH*

*University of Engineering and Technology, Vietnam National University, Hanoi, 144 Xuan Thuy Street, Cau Giay District, 10000 Ha Noi, Viet Nam*



**Abstract.** Reconstructing phylogenetic trees from protein sequences normally requires empirical amino acid substitution models to calculate the likelihood of trees or genetic distances between species. The tree of life is classified into three domains of Eukaryotes, Archaea, and Bacteria. The amino acid substitution models have been intensively studied for decades, but few are related to Bacteria. Rooting bacterial trees remains a challenging problem in the phylogenetic analysis due to the long branch separating Bacteria and other domains. The two main objectives of this paper are estimating amino acid substitution models Q.bac and NQ.bac for bacterial evolutionary studies and assessing the capability of the time non-reversible model NQ.bac in rooting bacterial trees. Experiments showed that both the time-reversible model (Q.bac) and the time-non-reversible model (NQ.bac) were significantly better than the existing models in analyzing bacterial protein sequences. Interestingly, the time non-reversible model NQ.bac helped reconstruct maximum likelihood bacterial trees with reliable roots for 177 (23.7%) out of 748 testing alignments without requiring predefined outgroups. This outgroup-free rooting method enhances the studies of bacterial evolution. We recommend researchers employ both Q.bac and NQ.bac models in analyzing bacterial protein sequences. The datasets and scripts used in this manuscript are available at https://doi.org/10.6084/m9.figshare.20457264.

**Keywords.** Amino acid substitution models, bacterial protein sequences, time-non-reversible models, time-reversible models.

## 1. INTRODUCTION

Amino acid (AA) substitution models are a powerful tool to study the relationships among species using their protein sequences. The likelihood or Bayesian tree construction methods require AA substitution models to calculate the likelihood of trees. The AA substitution models can be used as the score matrices to measure the distances between sequences in protein sequence similarity searches. Using improper AA substitution models might result in systematic errors in inferred trees [1] or Ancestral sequence construction [2].

The AA substitution model consists of a large number of parameters. Therefore, it must be estimated from an empirical dataset. A number of general models such as WAG [3], LG [4], Q.pfam [5], or NQ.pfam [6] have been estimated from large datasets including protein sequences from various species. A substitution model such as WAG or LG consists of a 20×20

*Corresponding author.
*E-mail addresses*: cuongdc@vnu.edu.vn (C.C. Dang); vinhls@vnu.edu.vn (L.S. Vinh).

matrix and a vector of 20 elements. The matrix represents the substitution rates between 20 amino acids while the vector contains the frequencies of 20 amino acids. The general models can be used for any protein sequences. However, they might not properly represent the evolutionary patterns of specific clades, so several clade-specific models have been estimated for plants, birds, etc. [5, 6]. Experiments showed that the clade-specific models are much better than the general models in analyzing their corresponding protein sequences.

The performance of models estimated from genome datasets using the maximum likelihood methods has been currently investigated based on simulation data [7]. The QMaker [5] and nQMaker [6] methods were examined in estimating both time-reversible and time-non-reversible amino acid substitution models from simulated genome datasets. The experiments showed that models estimated from genome datasets with greater than an equal to 100 genes highly correlated with the true models and performed well in reconstructing trees.

The short generation time and large population size of Bacteria help them evolve rapidly to adapt to environmental changes or immune responses from the hosts. This makes Bacteria the most diverse and abundant organisms on the Earth. Many bacterial protein sequences are available and appropriate for studying Bacteria because the rapid evolution of Bacteria might make their nucleotide sequences saturated. To date, there is no existence of amino acid substitution models specifically for analyzing bacterial protein sequences. We had to rely on the general amino acid substitution models, e.g., WAG [3] or LG [4], or clade-specific models such as Q.yeast or Q.insect [5]. The existing models are unlikely to be able to properly reflect the evolutionary patterns of Bacteria.

Rooting trees is an essential problem in phylogenetic analyses that has been studied for a long time [8]. There are several approaches to root phylogenetic trees, e.g., using an outgroup [8–10], assuming a molecular clock [11], or employing time non-reversible models of evolution [6,12,13]. Selecting outgroups from Archaea to root bacterial trees is controversial because the long distance from the outgroups to Bacteria might distort the tree structure of within-bacterial species. The molecular clock approach assumes a constant substitution rate along all lineages that might be biologically unrealistic especially when analyzing distantly related species because their substitution rates may vary during the long evolution process. Other rooting methods could be using gene duplication or indels [14, 15], minimal ancestor deviation [16], or minimum variance rooting [17]. Recently, a phylogenomic approach using information from gene duplications and losses within a genome together with gene transfers between genomes is proposed to root the tree without including an archaeal outgroup [18].

Using the time-non-reversible AA substitution model to reconstruct the maximum likelihood of rooted trees from protein alignments is a promising approach for Bacteria. In this paper, we collected 1748 bacterial protein alignments and used the maximum-likelihood methods to estimate both the time-reversible model (Q.bac) and the time-non-reversible model (NQ.bac) from 1000 alignments. We used the remaining 748 alignments to examine the performance of newly estimated bacterial models and the existing models. Experiments showed that bacterial models outperformed the existing models in analyzing bacterial protein sequences. The time-non-reversible model NQ.bac was better than the time-reversible models in a considerable number of cases. The time-non-reversible model NQ.bac helped reconstruct rooted bacterial trees of which 177 trees have root branches with support values $\geq 70\%$ indicating that these trees were correctly rooted with a high probability. To summarize, two main contributions of this paper are:

- Estimating two amino acid substitution models Q.bac and NQ.bac for bacterial evolutionary studies. Comparing the fit of the two new models to 31 existing models in reconstructing a bacterial tree.
- Assessing the capability of the time non-reversible model NQ.bac in rooting bacterial trees.

The rest of the paper is organized in three sections, Section 2 presents the proposed method, the workflow, and the evaluation metrics. Section 3 shows the experimental results and Section 4 is the conclusion of the paper.

## 2. MATERIALS AND METHODS

### 2.1. Data

To estimate Q.bac and NQ.bac, we obtained alignments from the HAMAP database [19]. HAMAP contains a collection of 2388 expert-curated protein families, and these protein families are used for protein family classification and functional annotation. Since we are interested in bacterial sequences only, we removed all non-bacterial alignments. We also removed duplicated sequences. The final bacterial protein dataset consists of 1748 alignments including 107842 sequences with a total of 711669 sites. On average, each alignment consists of 62 sequences with a length of 407 amino acids.

The alignments were randomly divided into two datasets: the training dataset including 1000 alignments and the testing dataset consisting of 748 remaining alignments. We estimated both time-reversible and time-non-reversible models from the training dataset and analyzed them against the 31 existing models on the 748 testing alignments.

### 2.2. Methods

We assume that substitutions among amino acids are independent among sites during the evolution and modeled by a time-homologous, time-continuous, and stationary Markov process. The substitution process can be expressed in terms of a matrix $Q = \{q_{xy}\}$ describing the instantaneous substitution rates between twenty amino acids. Precisely, the off-diagonal elements $q_{xy}$ $(x \neq y)$ represents the substitution rate from amino acid $x$ to amino acid $y$ while the diagonal elements $q_{xx}$ are calculated such that the sum of each row equals zero. In the phylogenetic tree, the branch length indicates the number of substitutions between two nodes, therefore, the matrix $Q$ is normalized so that the total number of substitutions per time unit is one. This effectively removes one parameter in the $Q$ matrix, i.e., the model consists of 379 parameters.

The amino acid substitution process can be assumed to be time-reversible (i.e., the exchangeability rates between two amino acids are the same in both directions) to simplify the model estimation process. The time-reversible property helps decompose the matrix $Q$ into a symmetric exchangeability rate matrix $R = \{r_{xy}\}$ and an amino acid frequency vector $\Pi = \{\pi_x\}$, i.e., $q_{xy} = \pi_y r_{xy}$ and $q_{xx} = -\sum_y q_{xy}$. As a result, the time-reversible model consists of 208 parameters much fewer than the time non-reversible model. Note that maximum likelihood trees reconstructed with time-reversible models are unrooted.

Let $\mathbf{A} = \{A_1, \ldots, A_n\}$ be a list of $n$ alignments. Let $\mathbf{T} = \{T_1, \ldots, T_n\}$ denote a list of trees for alignments $\mathbf{A}$. As we do not know the true tree for alignments $of$ $\mathbf{A}$, trees

determined by the maximum likelihood tree reconstruction method (e.g., IQ-TREE) can be used in the model estimation algorithms. In phylogenetic analyses, the rate heterogeneity among sites should be accounted for by a site rate model, e.g., Γ distribution of rates [20]. Let $\mathbf{H} = \{H_1, \ldots, H_n\}$ be the list of site rate models for alignments $\mathbf{A}$. The best-fit site rate model $H_i$ for alignment $A_i$ is normally selected by a model selection program, e.g., ModelFinder [21].

The maximum likelihood (ML) model estimation methods determine parameters of the substitution model $Q$ together with trees $\mathbf{T}$ and site rate models $\mathbf{H}$ to maximize the likelihood value $L(Q|\mathbf{T}, \mathbf{H}, \mathbf{A})$. Note that $L(Q|\mathbf{T}, \mathbf{H}, \mathbf{A})$ can be calculated from the likelihood values of alignments, i.e., $L(Q|\mathbf{T}, \mathbf{H}, \mathbf{A}) = \prod_{i=1\ldots n} L(Q|\mathbf{T}, \mathbf{H}, A_i)$. Optimizing $L(Q|\mathbf{T}, \mathbf{H}, \mathbf{A})$ from a set of alignments is a computationally complicated and expensive task.

Several approximate ML methods have been proposed to efficiently estimate AA substitution models [4–6, 22]. They discovered that the parameters of $Q$ can be efficiently estimated based on near-optimal trees $\mathbf{T}$ and site rate models $\mathbf{H}$. Thus, the parameters of $Q$, $\mathbf{T}$, and $\mathbf{H}$ can be estimated iteratively instead of simultaneously. The estimation process should be repeated several times until the parameters of $Q$ remain unchanged to optimize the likelihood value $L(Q|\ \mathbf{T},\ \mathbf{H}, \mathbf{A})$.

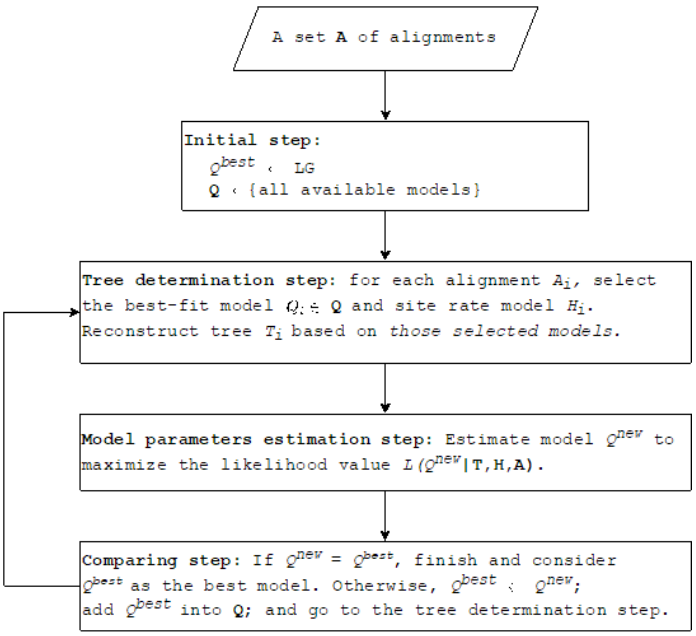


Figure 1: The flowchart to estimate amino acid substitution models for bacteria

The model estimation process is composed of four main steps (i.e., initial step, tree determination step, model estimation step, and comparing step) and is illustrated in Figure 1. The overall workflow is as follows:

1. *The initial step* assigns the general model LG [4] as the current best model, called $Q^{best}$, for the training alignments $\mathbf{A}$. Let $\mathbf{Q}$ be the set of existing amino acid substitution models.
2. The tree determination step builds trees for alignments of $\mathbf{A}$. For each alignment $A_i$, it selects the best-fit substitution model $Q_i \in \mathbf{Q}$ and the best-fit site rate model $H_i$

using the model selection program ModelFinder [21]. The ModelFinder program uses an initial parsimony tree to compute the likelihood value for each model in **Q**, and subsequently selects the best-fit substitution and site rate models that minimize the BIC score [23]. For each training alignment $A_i$, best-fit substitution model $Q_i$, and best-fit site rate model $H_i$, we infer one tree $T_i$ using the maximum likelihood tree construction package IQ-TREE 2 [24].

3. The model parameter optimization step determines a new matrix $Q^{new}$ to maximize the likelihood value $L\left(Q^{new}|\mathbf{T},\ \mathbf{H}\ ;\mathbf{A}\right)$. As $Q^{new}$ can be efficiently estimated with nearly optimal trees, we only re-estimate branch lengths of trees, but fix the tree topologies in this step to avoid computational burden. The tree topologies will be re-optimized if the new model $Q^{new}$ is considerably different from the current best model $Q^{best}$.
4. The comparing step uses the Pearson correlation to measure the similarity between $Q^{new}$ and $Q^{best}$. We flattened out two matrices into two 400-element vectors and computed their correlation. If the correlation is greater than 0.99, finish the estimation process and consider $Q^{best}$ as the final best model. Otherwise, assign $Q^{best}$ by $Q^{new}$, add $Q^{best}$ *to the model set* **Q**, and go to step 2. Experiments show that $Q^{best}$ is normally obtained after three iterations.

We applied the estimation pipeline (Figure 1) to estimate two new bacteria-specific models from 1000 bacterial protein alignments in the training dataset, i.e., using the QMaker method [5] to estimate the time reversible model Q.bac, and the nQMaker method [6] to estimate the time-non-reversible model NQ.bac. The current study on simulation data reveals that QMaker and nQMaker can estimate reliable time-reversible and time-non-reversible amino acid substitution models from genome datasets [7]. The authors simulated various genome datasets with different numbers of genes based on predefined models and predefined trees. The QMaker and nQMaker methods were used to estimate both time-reversible and time-non-reversible amino acid substitution models from the simulated genome datasets. The experiments showed that models estimated from simulated datasets with greater than and equal 100 genes were highly correlated with the true models (Pearson correlations $\geq$ 0.995), and performed well in reconstructing trees.

In the maximum likelihood phylogenetic analyses, the model $Q$ (e.g., Q.bac or NQ.bac) is used to calculate the likelihood value $L\left(T|Q,H,A\right)$ of a tree $T$ given the model $Q$, the site-rate model $H$, and the alignment $A$ as followings

$$L\left(T|Q,\ H,\ A\right)=\prod_{j=1}^{l}L\left(T|Q,\ H,\ A^{j}\right)=\prod_{j=1}^{l}\text{Prob}\left(A^{j}|T,\ Q,\ H\right), \tag{1}$$

where, $l$ is the length of alignment $A$, $A^j$ is the data at site $j$ of alignment $A$, $L\left(T|Q,\ H,\ A^{j}\right)$ is the likelihood value of tree $T$ at site $j$ that can be computed by the conditional probability $\text{Prob}\left(A^{j}|T,\ Q,\ H\right)$ of data $A^j$. Technically, the matrix $Q$ is used to calculate the transition probability matrix $P(t)=e^{Qt}$ (i.e., $P_{xy}(t)$ is the probability of changing from amino acid $x$ to amino acid $y$ after $t$ time units) in computing the conditional probability $\text{Prob}\left(A^{j}|T,\ Q,\ H\right)$ using the pruning algorithm [25].

We examined the performance of Q.bac, NQ.bac, and the 31 available reversible and non-reversible models (see supplementary Table 3) on the 748 testing alignments using the Bayesian information criterion (BIC) [23]. The BIC score combines both likelihood value

$L(Q|\ \mathbf{T},\ \mathbf{H}, \mathbf{A})$ and the number of parameters in the models used to build the trees to assess the performance. A previous study showed that the BIC criteria and the AIC criteria [26] gave similar results [27]. Therefore, in this paper, we reported and discussed the results using the BIC scores.

We utilized a resampling estimated log likelihoods (RELL) test [28] with 10000 bootstraps and a $p_{RELL}$ threshold of 0.05 to calculate the number of alignments that the Q.bac model is significantly better than the NQ.bac model and vice versa. Consider an alignment $A$ with $l$ sites, we computed the likelihood values for every site of $A$ using the NQ.bac and Q.bac models. If BIC value ($BIC_{Q.bac}$) of the tree inferred with Q.bac is smaller than the BIC value ($BIC_{NQ.bac}$) of the tree inferred with NQ.bac (i.e., $\Delta BIC_{original} = BIC_{NQ.bac} - BIC_{Q.bac} > 0$), we applied the RELL procedure to create 10000 replicates of sites (i.e., each replicate consists of $l$ sites randomly sampled with replacement from the sites of $A$). For each replicate of sites, the BIC value ($REP_{Q.bac}$) with the Q.bac model and the BIC value ($REP_{NQ.bac}$) with the NQ.bac model were calculated to determine the BIC difference (i.e., $\Delta BIC_{replicate} = REP_{NQ.bac} - REP_{Q.bac}$) between two models on the replicate of sites. The $p_{RELL}$ was calculated as the percentage of replicates that had $\Delta BIC_{replicate} > 2\ \times \Delta BIC_{original}$.

To evaluate the impact of the models on tree structures, we calculated the normalized robinson and foulds (nRF) distance [29] between trees reconstructed with different models. The nRF distance between two trees is the unshared splits between them divided by the total number of all splits. The nRF distance ranges from 0 (two identical trees) to 1 (two completely different trees).

Since the time non-reversible model NQ.bac allows us to reconstruct rooted trees for bacterial alignments, we used the time non-reversible model NQ.bac to reconstruct a rooted tree $T_r$ for each testing alignment; each branch $e$ on $T_r$ is labeled with a rootstrap value [13] indicating the probability that the root is placed in the branch $e$. The rootstrap value of a branch is calculated as the fraction of 1000 rooted bootstrap trees which have the root on that branch. The 1000 bootstrap trees were produced by using the ultrafast bootstrap analysis [1].

Additionally, we performed the approximately unbiased (AU) test [30] with 10000 replicates. The AU test compares the log-likelihoods of the trees being re-rooted on each branch of $T_r$. The branches with $p_{AU} < 0.05$ will be rejected as the root placement [13].

## 3. RESULTS

### 3.1. Model performance

First, we examined the fit of the 31 existing models on bacteria alignments by comparing the Bayesian information criterion scores (BIC) [23] of ML trees constructed with these models on the 748 testing alignments. Since examined models have different numbers of free parameters (reversible models such as LG and Q.pfam have one free parameter while non-reversible models such as NQ.pfam have two free parameters), the BIC scores were used instead of the original likelihood value. This test would propose a list of current top models best fit with bacteria alignments and we will eventually compare the fit of Q.bac, NQ.bac, and the four top models. All ML trees were constructed with the site rate model I+$\Gamma$4 (*Gamma* distribution with four categories and one category of invariant). We found that the general models NQ.pfam, LG, Q.pfam, and WAG were the best-fit substitution models for 240, 234,

91, and 34 testing alignments, respectively. The clade-specific models Q.yeast, Q.plant, and Q.insect were the best substitution models for some testing alignments (i.e., Q.yeast for 47 alignments, Q.plant for 16 alignments, and Q.insect for 14 alignments). Among the existing models, the general models such as NQ.pfam and LG help build better maximum likelihood bacterial trees than the other models. A majority of bacterial alignments yield better BIC scores when constructing ML trees with the general time-non-reversible NQ.pfam and the general time-reversible LG substitution models.

Table 1 presents the average BIC values per site of Q.bac, NQ.bac, and four top existing models on 748 testing alignments. The results show that NQ.bac and Q.bac are the best-fit models for analyzing bacterial alignments. The WAG model is worse than the other models (e.g., the average BIC/site of WAG is 0.37 lower than that of LG). Two models, Q.pfam and NQ.pfam are slightly better than LG with a BIC/site improvement of 0.06 and 0.09, respectively.

Table 1: The average BIC values per site of six main models on 748 bacteria testing alignments. Note: All models were tested with four categories of gamma-distributed and one category of invariant rates (I+Γ4).

| Model | BIC/site | BIC/site gain compared to WAG |
|---|---|---|
| WAG | 101.82 | 0 |
| LG | 101.45 | 0.37 |
| Q.pfam | 101.39 | 0.43 |
| NQ.pfam | 101.35 | 0.46 |
| Q.bac | 101.30 | 0.52 |
| NQ.bac | 101.26 | 0.56 |

Second, we compared directly the fit of Q.bac and NQ.bac on the 748 testing alignments using the BIC scores. The time-non-reversible model NQ.bac was better than the time-reversible model Q.bac, i.e., NQ.bac was better (worse) than Q.bac on 465 (283) out of 748 testing alignments.

Figure 2 summarizes the comparisons between Q.bac and NQ.bac with respect to alignment sizes (i.e., the number of sequences and the number of sites). The NQ.bac model was significantly better than the Q.bac model on alignments with more than 100 sequences (Figure 2a) or 600 sites (Figure 2b). The explanation is that the time-non-reversible model consists of more parameters, so it fits well with large alignments.

Finally, we assessed the performance (BIC scores) of Q.bac, NQ.bac, and 31 available reversible and non-reversible models (see supplementary Table 3) on 748 testing alignments.

Figure 3 shows that the NQ.bac and Q.bac models outperformed all other models on 252 (33.7%) and 172 (23.0%) testing alignments, respectively. The general reversible model LG was the best model for 84 (11.2%) testing alignments. The Q.yeast model estimated from yeasts was the best-fit model for 60 alignments (about 8%). The other models were selected as the best models for several alignments. These results indicate the advantage of newly estimated bacterial models over the existing models in analyzing bacterial protein data.

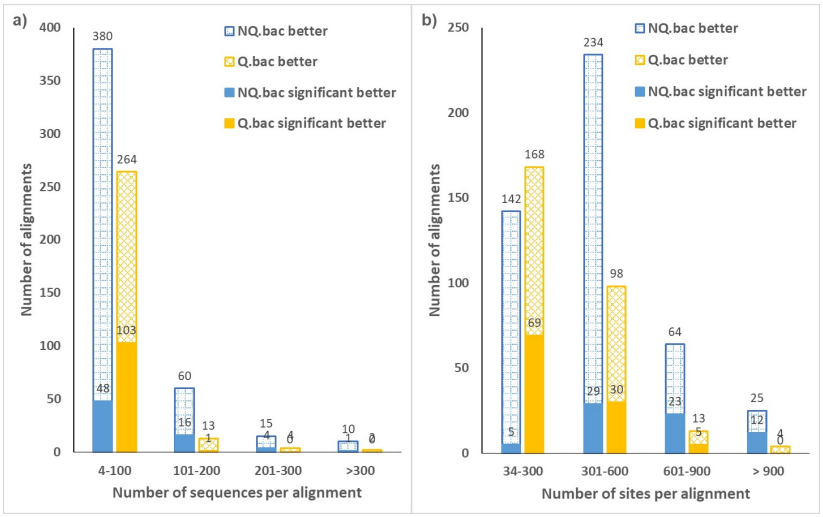


Figure 2: Comparisons between the time-reversible model Q.bac and the time-non-reversible model NQ.bac on 748 test alignments. The NQ.bac model was significantly better than the Q.bac model on alignments having more than 100 sequences (a) or 600 sites (b).

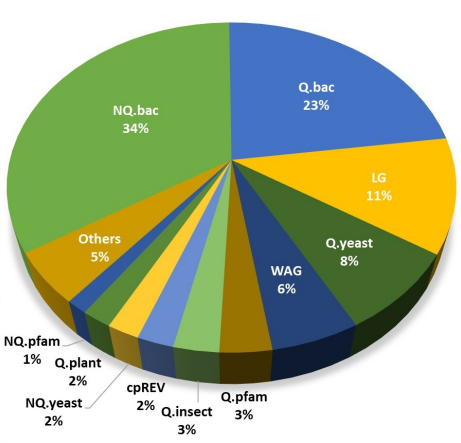


Figure 3: The percentage of testing alignments that each amino acid substitution model was selected as the best-fit model.

### 3.2. Model analysis

We used the principal component analysis (PCA) to visualize the differences among models. Each model was represented by one vector of 400 amino acid substitution rates in the $Q$ matrices. The objective of this PCA analysis is to show how much the new models Q.bac and NQ.bac are similar to or different from existing models. Figure 4 illustrates the PCA analysis of Q.bac, NQ.bac, and the existing models excluding models for mtDNA-proteins and viruses because they are not highly correlated with the bacteria models. The models are separated into two sides: the time-reversible models are on the left side and the time-non-reversible models are on the right side. The Q.bac, Q.insect, and Q.yeast are grouped together with general time-reversible models (e.g., WAG, LG, and Q.pfam) except JTT model that was estimated a long time ago from a small dataset. These models are

far away from clade-specific models for plants and animals (i.e., Q.plant, Q.mammal, and Q.bird). The time-non-reversible model NQ.bac is not close to other time-non-reversible models. The NQ.insect and NQ.yeast are the two closest models to NQ.bac. Interestingly, the time-non-reversible model NQ.pfam is close to the NQ.bird, but far away from other clade-specific time-non-reversible models (i.e., NQ.mammal, NQ.plant, NQ.insect, NQ.yeast, and NQ.bac). In other words, the general time-non-reversible NQ.pfam might not be suitable for analyzing clade-specific data.

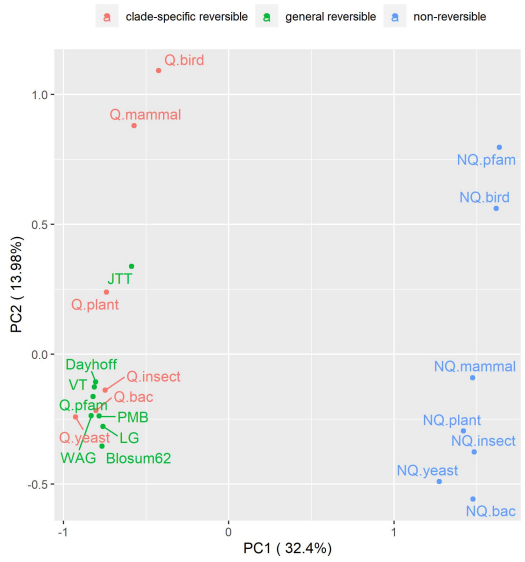


Figure 4: Principal component analysis (PCA) of NQ.bac, Q.bac, and the existing models (except models for mtDNA-proteins and viruses) based on their amino acid substitution rates.

We also directly compared the substitution rates of Q.bac with those of NQ.bac, Q.yeast, LG, and NQ.pfam to emphasize the differences between Q.bac, NQ.bac, and other models. Notable differences between these models were identified and illustrated in Figure 5. For example, 65 (104) substitution rates in Q.bac are at least two times smaller (larger) than those in NQ.bac. Similarly, 75 (91) substitution rates in Q.bac are at least twice small (large) than ones in the Q.yeast.

### 3.3. Topological impacts

The nRF distances between trees constructed with the time- reversible model Q.bac and the time-non-reversible model NQ.bac is 0.1 (see detail in Table 2). Thus, on average, about 10% of splits appearing in one tree constructed with Q.bac will not appear in the corresponding tree constructed with NQ.bac. The average nRF distances between 748 trees constructed with Q.bac and those with general time-reversible models WAG, LG, and Q.pfam are 0.12, 0.09, and 0.08, respectively. Similarly, the average nRF distances between trees constructed with the time-non-reversible model NQ.bac and those with the time-non-reversible model NQ.pfam is 0.09. These results indicate that the amino acid substitution models have a considerable impact on building tree topologies for bacterial alignments.

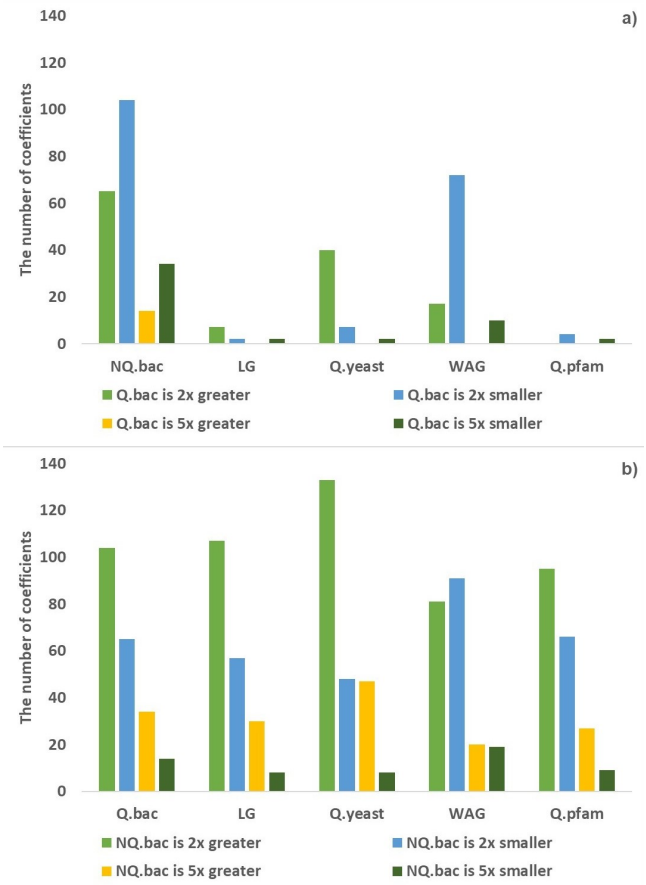


Figure 5: The comparison between the coefficients of Q.bac (a), NQ.bac (b) with NQ.bac (Q.bac), LG, Q.yeast and NQ.pfam models. Notations: 2x (5x) indicates two (five) times difference.

Table 2: The average nRF distances between 748 trees constructed with six models NQ.bac, Q.bac, NQ.pfam, Q.pfam, LG, and WAG.

| Model | NQ.bac | Q.bac | NQ.pfam | Q.pfam | LG |
|---|---|---|---|---|---|
| Q.bac | 0.10 | | | | |
| NQ.pfam | 0.09 | 0.10 | | | |
| Q.pfam | 0.10 | 0.08 | 0.09 | | |
| LG | 0.11 | 0.09 | 0.10 | 0.08 | |
| WAG | 0.12 | 0.12 | 0.12 | 0.11 | 0.12 |

### 3.4. Building rooted tree

Figure 6 shows that 361 (48%) out of 748 rooted trees reconstructed from testing alignments with the NQ.bac model have root branches with rootstrap values greater than 50% and confirmed by the AU test (there is only one rooted tree whose root branch with rootstrap value greater than 50% but failed the AU test). Among these, 177 rooted trees have

considerably reliable root branches with rootstrap values greater than or equal to 70%. Note that 90 of the considerably reliable rooted trees were fit best with the NQ.bac model. The results showed the capability of the time-non-reversible model NQ.bac in rooting bacterial trees.

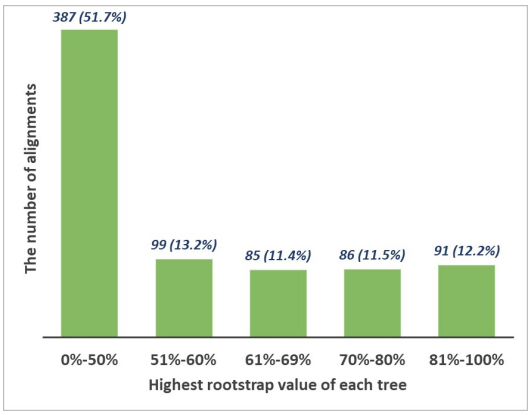


Figure 6: The distribution of rooted trees with the highest rootstrap values

## 4. CONCLUSIONS

The rapid evolution of Bacteria might make their nucleotide sequences saturated, so the data of protein sequences is a proper alternative data source to study their evolutionary relationships. To this end, amino acid substitution models must be specifically estimated for Bacteria because using the existing amino acid substitution models constructed from Eukaryotes or viruses might lead to systematic errors in analyzing bacterial protein sequences.

We have estimated both the time-reversible model (Q.bac) and the time-non-reversible model (NQ.bac) from a thousand bacterial alignments. Experiments showed that the bacterial models outperformed the existing models in reconstructing maximum likelihood trees for bacterial species. The NQ.bac model consists of more parameters than the Q.bac model, therefore, it is more suitable for analyzing large alignments than small alignments. Users should use a model selection program such as ModelFinder [21] to select the best-fit model for their alignments under the study.

To date, using outgroups is a widely used approach to root phylogenetic trees. However, determining proper outgroups to root bacterial trees is problematic because of the long distance between Bacteria and other kingdoms. The long branch between the outgroups and other species in the bacterial trees might twist topological structures among bacterial species. An alternative approach to building rooted trees is using the maximum likelihood tree reconstruction methods with time-non-reversible models.

In this paper, we used the time-non-reversible model NQ.bac to reconstruct rooted trees for Bacteria. Nearly half of the trees have root branches with support values greater than 50% and verified by the approximately unbiased statistical test. Among these trees, half of them have root branches with high rootstrap values of greater or equal to 70%. These results indicate that rooting trees by using the time-non-reversible amino acid substitution models is a plausible and promising approach for Bacteria.

This paper has two main contributions, and they meet our proposed objectives. Firstly, it introduces two new amino acid substitution models Q.bac and NQ.bac for bacterial evolutionary studies. The experiments show that the fit of the two new models is significantly better than that of 31 existing models in reconstructing bacterial trees. Secondly, NQ.bac helps reconstruct maximum likelihood bacterial trees with reliable roots without requiring predefined outgroups for 177 (23.7%) out of 748 testing alignments.

## REFERENCES

[1] D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh, "UFBoot2: Improving the ultrafast bootstrap approximation," *Molecular Biology and Evolution*, vol. 35, pp. 518–522, Oct. 2017, Doi: 10.1093/molbev/msx281.

[2] R. Del Amparo and M. Arenas, "Consequences of substitution model selection on protein ancestral sequence reconstruction," *Molecular Biology and Evolution*, vol. 39, Jul. 2022, Doi: 10.1093/molbev/msac144.

[3] S. Whelan and N. 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, May 2001, Doi: 10.1093/oxfordjournals.molbev.a003851.

[4] S. Q. Le and O. Gascuel, "An improved general amino acid replacement matrix," *Molecular Biology and Evolution*, vol. 25, no. 7, 2008, Doi: 10.1093/molbev/msn067.

[5] B. Q. Minh, C. C. Dang, L. S. Vinh, and R. Lanfear, "QMaker: Fast and accurate method to estimate empirical models of protein evolution," *Systematic Biology*, vol. 70, no. 5, pp. 1046–1060, May 2021, Doi: 10.1093/sysbio/syab010.

[6] C. C. Dang et al., "nQMaker: Estimating time non-reversible amino acid substitution models," *Systematic Biology*, 2022, Doi: 10.1101/2021.10.18.464754.

[7] N. H. Tinh, C. C. Dang, and L. S. Vinh, "Estimating amino acid substitution models from genome datasets: A simulation study on the performance of estimated models," *Journal of Evolutionary Biology*, vol. 37, no. 2, pp. 256–265, 2024, Doi: https://doi.org/10.1093/jeb/voad017.

[8] W. P. Maddison, M. J. Donoghue, and D. R. Maddison, "Outgroup analysis and parsimony," *Systematic Biology*, vol. 33, no. 1, 1984, Doi: 10.1093/sysbio/33.1.83.

[9] Z. Yang and D. Roberts, "On the use of nucleic acid sequences to infer early branchings in the tree of life," *Molecular Biology and Evolution*, vol. 12, no. 3, pp. 451–458, May 1995, Doi: 10.1093/oxfordjournals.molbev.a040220.

[10] J. P. Huelsenbeck, J. P. Bollback, and A. M. Levine, "Inferring the root of a phylogenetic tree," *Systematic Biology*, vol. 51, no. 1, 2002, Doi: 10.1080/106351502753475862.

[11] S. Y. W. Ho and S. Duchêne, "Molecular-clock methods for estimating evolutionary rates and timescales," *Molecular Ecology*, vol. 23, no. 24, pp. 5947–5965, 2014, Doi: https://doi.org/10.1111/mec.12953.

[12] B. Bettisworth and A. Stamatakis, "Root digger: A root placement program for phylogenetic trees," *BMC Bioinformatics*, vol. 22, no. 1, 2021, Doi: 10.1186/s12859-021-03956-5.

[13] S. Naser-Khdour, B. Quang Minh, and R. Lanfear, "Assessing confidence in root placement on phylogenies: An empirical study using nonreversible models for mammals," *Systematic Biology*, vol. 71, no. 4, 2022, Doi: 10.1093/sysbio/syab067.

[14] N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata, "Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 23, 1989, Doi: 10.1073/pnas.86.23.9355.

[15] J. A. Lake, C. W. Herbold, M. C. Rivera, J. A. Servin, and R. G. Skophammer, "Rooting the tree of life using nonubiquitous genes," *Molecular Biology and Evolution*, vol. 24, no. 1, 2007, Doi: 10.1093/molbev/msl140.

[16] F. D. K. Tria, G. Landan, and T. Dagan, "Phylogenetic rooting using minimal ancestor deviation," *Nature Ecology & Evolution*, vol. 1, p. 193, 2017, Doi: 10.1038/s41559-017-0193.

[17] U. Mai, E. Sayyari, and S. Mirarab, "Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction," *PLOS ONE*, vol. 12, no. 8, p. e0182238, 2017, Doi: 10.1371/journal.pone.0182238

[18] G. A. Coleman et al., "A rooted phylogeny resolves early bacterial evolution," *Science*, vol. 372, p. eabe0511, 2021, Doi: 10.1126/science.abe0511.

[19] T. Lima et al., "HAMAP: A database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot," *Nucleic Acids Research*, vol. 37, pp. D471–D478, Oct. 2008, Doi: 10.1093/nar/gkn661.

[20] Z. Yang, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites," *Molecular Biology and Evolution*, vol. 10, no. 6, 1993, Doi: 10.1093/oxfordjournals.molbev.a040082.

[21] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin, "ModelFinder: Fast model selection for accurate phylogenetic estimates," *Nature Methods*, vol. 14, pp. 587–589, 2017, Doi: 10.1038/nmeth.4285.

[22] C. C. Dang, V. S. Le, O. Gascuel, B. Hazes, and Q. S. Le, "FastMG: A simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets," *BMC Bioinformatics*, vol. 15, 2014, Doi: 10.1186/1471-2105-15-341.

[23] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461 – 464, 1978, Doi: 10.1214/aos/1176344136.

[24] B. Q. Minh et al., "IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era," *Molecular Biology and Evolution*, vol. 37, no. 5, 2020, Doi: 10.1093/molbev/msaa015.

[25] J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376, 1981, Doi: 10.1007/BF01734359.

[26] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974, Doi: 10.1109/TAC.1974.1100705.

[27] V. S. Le, C. C. Dang, and Q. S. Le, "Improved mitochondrial amino acid substitution models for metazoan evolutionary studies," *BMC Evolutionary Biology*, vol. 17, p. 136, 2017, Doi: 10.1186/s12862-017-0987-y.

[28] H. Kishino and M. Hasegawa, "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea," *Journal of Molecular Evolution*, vol. 29, no. 2, pp. 170–179, Aug. 1989.

[29] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1, pp. 131–147, 1981, Doi: https://doi.org/10.1016/0025-5564(81)90043-2.

[30] H. Shimodaira, "An approximately unbiased test of phylogenetic tree selection," *Systematic Biology*, vol. 51, no. 3, 2002, Doi: 10.1080/10635150290069913.

*Received on October 31, 2023*
*Accepted on January 26, 2024*

## APPENDIX

Table 3: Existing amino acid substitution models

| | **Matrix** | **Genomic regions** | **Time non-reversible or not** |
|---|---|---|---|
| 1 | Blosum62 | General | Time reversible |
| 2 | Dayhoff | General | Time reversible |
| 3 | JTT | General | Time reversible |
| 4 | LG | General | Time reversible |
| 5 | PMB | General | Time reversible |
| 6 | VT | General | Time reversible |
| 7 | WAG | General | Time reversible |
| 8 | mtArt | Mitochondrial | Time reversible |
| 9 | mtMam | Mitochondrial | Time reversible |
| 10 | mtRev | Mitochondrial | Time reversible |
| 11 | mtZoa | Mitochondrial | Time reversible |
| 12 | mtMet | Mitochondrial | Time reversible |
| 13 | mtVer | Mitochondrial | Time reversible |
| 14 | mtInv | Mitochondrial | Time reversible |
| 15 | HIVb | Viral | Time reversible |
| 16 | HIVw | Viral | Time reversible |
| 17 | FLU | Viral | Time reversible |
| 18 | FLAVI | Viral | Time reversible |
| 19 | rtREV | Viral | Time reversible |
| 20 | Q.bird | General | Time reversible |
| 21 | Q.insect | General | Time reversible |
| 22 | Q.mammal | General | Time reversible |
| 23 | Q.pfam | General | Time reversible |
| 24 | Q.plant | General | Time reversible |
| 25 | Q.yeast | General | Time reversible |
| 26 | NQ.bird | General | Time non-reversible |
| 27 | NQ.insect | General | Time non-reversible |
| 28 | NQ.mammal | General | Time non-reversible |
| 29 | NQ.pfam | General | Time non-reversible |
| 30 | NQ.plant | General | Time non-reversible |
| 31 | NQ.yeast | General | Time non-reversible |