Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improving the study of plant evolution with multi-matrix mixture models

Tinh Nguyen Huy
University of Engineering and Technology, Vietnam National University
Vinh Le Sy
vinhls@vnu.edu.vn

University of Engineering and Technology, Vietnam National University

Short Report

Keywords: Plant evolution, Maximum likelihood estimation method, amino acid substitution models, time reversible models, mixture models

Posted Date: November 17th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3617795/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Plant Systematics and Evolution on April 12th, 2024. See the published version at https://doi.org/10.1007/s00606-024-01896-0.

Abstract

Amino acid substitution model is a key component to study the plant evolution from protein sequences. Although single-matrix amino acid substitution models have been estimated for plants (i.e., Q.plant and NQ.plant), they are not able to describe the rate heterogeneity among sites. A number of multi-matrix mixture models have been proposed to handle the site-rate heterogeneity, however, none are specifically estimated for plants. To enhance the study of plant evolution, we estimated both time reversible and time non-reversible multi-matrix mixture models QPlant.mix and NQPlant.mix from the plant genomes. Experiments showed that the new mixture models were much better than the existing models for plant alignments. We recommend researchers to use the new mixture models for studying the plant evolution.

Introduction

Modeling amino acid substitution process is one of the main problems in bioinformatics. In maximum likelihood approach, the models are used to compute substitution probabilities along branches of phylogenetic trees. Usually, this process is described by a Hidden Markov Model with time-continuous, time-homogeneous and stationary properties (Richard Durbin 2006). The central part of the model is a matrix of 20×20substitution rates each coefficient represents the substitution rate from amino acid to another amino acid. To reduce the computational complexity, the process might be assumed to be time-reversible, i.e., the exchangeability rates between two amino acids are the same at both directions.

A number of general models have been introduced such as JTT, WAG or LG (Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008). Recently, we introduced clade-specific models for some important clades including plants (Minh et al. 2021; Dang et al. 2022). They are single-matrix models each has one matrix describing the substitution rates among amino acids for all sites. It is well-known that the evolution rate is heterogeneous among sites and be affected by many factors like solvent accessibility and protein structure (Le and Gascuel 2010; Le et al. 2012). The site rate heterogeneity can be modeled by a discrete gamma distribution (Yang 1993), however, that is not a comprehensive solution from biological perspectives.

The multi-matrix mixture models have been proposed to handle the site-rate heterogeneity problem, e.g., LG4X, LG4M (Le et al. 2012). Although these mixture models outperform the single-matrix models, they were estimated from general datasets including diverse species and might be not appropriate for analyzing specific clades. To enhance the evolutionary studies of the plants, we estimated two new multi-matrix mixture models for plant species: time reversible model QPlant.mix and time non-reversible model NQPlant.mix. We examined the performance and impacts of the new mixture models with other existing models on plant alignments.

Materials and methods

Data

We employed the Plant genome dataset (Ran et al. 2018) that was used to estimate Q.plant (Minh et al. 2021) and NQ.plant (Dang et al. 2022). The dataset has 1308 alignments containing about 16,416,532 million amino acids. In average, each alignment consists of about 38 sequences and 330 sites. The dataset was divided into two parts: the training part of 1000 alignments and the testing part of 308 remaining alignments.

Methods

Single-matrix model

The substitution rates among amino acids are described by a single replacement matrix $Q = \{q_{xy}\}$ of 20×20elements where q_{xy} is the substitution rate from amino acid x to amino acid $y(x \neq y)$. If the substitution rates among amino acids are assumed to be time reversible (i.e., the exchangeability rates between two amino acids are the same in both directions), the matrix Q can be decomposed into two parts: a symmetric exchangeability rate matrix $R = \{r_{xy}\}$ and an amino acid frequency vector $\Pi = \{\pi_x\}$ such that $q_{xy} = \pi_y r_{xy}$ and $q_{xx} = -\sum_y q_{xy}$. The amino acid substitution processes among sites are assumed to be independent. The likelihood of a tree T and a replacement matrix Q given an alignment $D = (D_1, D_2, \ldots, D_l)$ of l sites is calculated from the likelihood of sites:

$$L(Q, T|D) = \prod_{i=1\ldots l} L(Q, T|D_i)$$

1

where $L(Q, T|D_i)$ is the likelihood of T and Q given the data $D_i$at site i.

Site rate heterogeneity

It is well-known that the substitution rates among sites are different (Yang 1993; Lartillot and Philippe 2004; Le et al. 2008, 2012; Quang et al. 2008; Wang et al. 2008). The site rate heterogeneity can be simply modeled by a discrete gamma distribution with K equally weighted rate categories (Yang 1993). With $K = 4$, the likelihood $L(Q, T|D)$ is calculated as followings:

$$L(Q, T, \alpha|D) = \prod_i \left(\frac{1}{4} \sum_{k=1}^{4} L(T, \Gamma(\alpha, k)Q|D_i) \right)$$

2

where $\Gamma(\alpha, r)$ is the k^{th} rate of the discrete gamma distribution with shape parameter α (the weights of rate categories are all equal to).

Multi-matrix model

The single-matrix model assumes that the substitution rates among amino acids are the same for all rate categories. To relax the unrealistic biological assumption, multi-matrix models LG4M and LG4X have been proposed (Le et al. 2012). The models have four replacement matrices for four different rate categories. Technically, the 4-matrix model **Q** includes four replacement matrices: $Q_1, Q_2, Q_3, \text{and} Q_4$ for 'very slow', 'slow', 'medium', and 'fast' rate category, respectively. The likelihood $L(\varvec{Q}, T|D)$ is now calculated as followings:

$$L\left(\mathbf{Q}=\{Q_1,Q_2,Q_3,Q_4\},W=\{w_1,w_2,w_3,w_4\},T|D\right)=\prod_i\left(\sum_{k=1}^{4}w_kL\left(T,Q_k|D_i\right)\right)$$

3

where w_k is the weight of matrix Q_k with a constraint $\sum_{k=1}^{4} w_k = 1$.

The LG4M model assigns each replacement matrix to one rate category obtained from the 4-category discrete gamma distribution with equal weights for all categories; while The LG4X model uses the distribution-free scheme that assigns different weights and rates for the rate categories (Le et al. 2012).

The maximum likelihood estimation method

The training dataset used to estimate an amino acid substitution model includes N alignments denoted by $\mathbf{D}=\{D^1,\ldots,D^N\}$. Let $\mathbf{T}=\{T^1,\ldots,T^N\}$ be the set of *N* trees corresponding to the *N* training alignments, i.e., T^i is the phylogenetic tree of alignment D^i. To account the site rate heterogeneity in the model estimation process, let $\mathrm{P}=\{\mathrm{P}^1,\ldots,\mathrm{P}^N\}$ and $\mathbf{W}=\left\{W^1,\ldots,W^N\right\}$ be the set of rates and weights of the training alignments, respectively (i.e., P^i and W^i are the rates and weights of alignment D^i). The 4-matrix model $\mathbf{Q}^*=\{Q_1^*,Q_2^*,Q_3^*,Q_4^*\}$ will be optimized to maximize the likelihood $L(\varvec{T}, \varvec{Q}, \varvec{P}, \varvec{W}|\varvec{D})$:

$$\mathbf{Q}^*=\{Q_1^*,Q_2^*,Q_3^*,Q_4^*\}=argmax_{\mathbf{Q}=\{Q_1,Q_2,Q_3,Q_4\}}\left\{\prod_{a=1\ldots N}L\left(T^a,\mathbf{Q},\mathrm{P}^a,W^a|D^a\right)\right\}.$$

4

Optimizing **Q**, **T**, **P**, and **W** simultaneously to determine the optimal **Q*** is computationally infeasible for large datasets. To overcome the computational burden, **Q**, **T**, **P**, and **W** are iteratively estimated to obtained **Q*** (Whelan and Goldman 2001; Le and Gascuel 2008; Le et al. 2008). In this paper, we enhanced the estimation process of (Le et al. 2012) by using IQ-TREE (Minh et al. 2020) to estimate trees and site rate models; QMaker (Minh et al. 2021) to estimate parameters of time reversible replacement matrices; and nQMaker (Dang et al. 2022) to estimate parameters of time non-reversible substitution rate matrices.

We estimated both time reversible model QPlant.mix and time non-reversible model NQPlant.mix from the 1000 training plant alignments. Each model contains 4 matrices corresponding to “very slow”, “slow”, “medium” and “fast” rate categories. Since the mixture models with the free-scheme rate distribution are better than that with the discrete gamma distribution (Le and Gascuel 2008; Minh et al. 2021; Dang et al. 2022), both QPlant.mix and NQPlant.mix were estimated with the free-scheme rate distribution.

Model performance assessment

We compared the performance models in building maximum likelihood trees including two new mixture models QPlant.mix and NQPlant.mix; the general single-matrix models LG, WAG, JTT and Q.pfam (Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008; Minh et al. 2021); the current single-matrix models Q.plant and NQ.plant for plants (Minh et al. 2021; Dang et al. 2022); and the general 4-matrix mixture models LG4X and LG4M (Le et al. 2012). Because the number of free parameters of the models are different, we used the BIC criterion (Schwarz 2007) to compare their performance.

We also examined the impacts of the models in building tree topologies. The topological difference between two trees constructed with two different models is measured by the Robinson and Foulds (RF) distance (Robinson and Foulds 1981). The RF distance is very wildly used metric which calculates the number of clades that belong to one tree but not to the other. The normalized nRD distance calculated by dividing to the total number of clades ranges from 0 (i.e., two trees have identical topologies) to 1 (i.e., the topologies of two trees are completely different).

Results

Model analysis

First, we calculated the similarity between time reversible models Q.plant and QPlant.mix; and time non-reversible models NQ.plant and QPlant.mix. Table 1 shows the Pearson correlations between matrices of the models as well as the relative difference between substitution rates among amino acids. We observe high Pearson correlations between the matrices of the mixture models with that of the single models. There are a number of substitution rates in mixture models that are at least two or five times different from that in the single-models, however, their squared errors are low (see Fig. 1). The mean squared errors of “very slow”, “slow”, “medium” and “fast” matrices of QPlant.mix compared to Q.plant are 0.0152, 0.0046, 0.0024, 0.0039, respectively. The mean squared errors between NQPlant.mix and NQ.plant are 0.0053 for “very slow”, 0.0021 for “slow”, 0.0015 for “medium”, and 0.0041 for “fast” rate category.

Table 1
The correlations between matrices of NQplant.mix (QPlant.mix) and that of NQ.plant (Q.plant). 2x (-2x) means the number of exchangeability coefficients of NQPlant.mix (or QPlant.mix) that are two times bigger (smaller) than that of NQ.plant (or Q.plant). Similar meanings for 5x and − 5x.

		Very slow	Slow	Medium	Fast
NQPlant.mix & NQ.plant	Correlation	0.971	0.983	0.988	0.971
	2x	71	21	11	58
	5x	23	4	0	12
	-2x	107	163	104	119
	-5x	78	74	67	86
QPlant.mix & Q.plant	Correlation	0.956	0.972	0.984	0.974
	2x	34	16	19	42
	5x	5	4	1	3
	-2x	154	194	92	99
	-5x	66	76	44	50

We also used the principal component analysis (PCA) to visualize the overall difference between matrices (see Fig. 2). The PCA shows that NQPlant.mix, QPlant.mix and 11 other models are grouped into several clusters. For example, the Q.plant, NQ.plant and the “very slow”, “slow” and “medium” matrices of QPlant.mix and NQPlant.mix are close to each other. The new mixture models for plants are far away from the general mixture models LG4X and LG4M.

Fitness and Topology comparison

We tested the performance of different models in building maximum likelihood trees on 308 plant testing alignments (see Fig. 3) using the BIC criterion. The two new mixture models for plants outperformed the other models tested on plant alignments. QPlant.mix was better than Q.plant on 301 over 308 alignments. Similarly, NQPlant.mix outperformed NQ.plant on 301 alignments. The QPlant.mix and NQPlant.mix were better than the general mixture models LG4X and LG4M, e.g., QPlant.mix fits better than LG4X on 304 alignments. We note that the mixture models QPlant.mix and NQPlant.mix were much better than the single models.

We also investigated the impacts of the models in building tree topologies. We calculated the nRF distances between trees constructed with our new models and that with the other models (see Table 2). The results show that the models considerably affect the tree topology. There are only a small percentage of testing alignments that trees inferred with the new mixture models and other models have the same topology. For example, QPlant.mix and Q.plant resulted in the same tree topology for only 52 out of 308

alignments. The average nRF distances between trees constructed with the new mixture models and that with the other models range from 0.099 to 0.142.

Table 2
The nRF distances between trees constructed with the new mixture models and that with the other models on testing alignments. $\#nRF = 0$: two trees have the same topology. Avg(nRF): the average nRF distances.

Model	QPlant.mix		NQPlant.mix	
	#nRF = 0	avg(nRF)	#RF = 0	avg(nRF)
LG	30	0.132	15	0.135
WAG	13	0.136	13	0.137
JTT	30	0.117	28	0.115
Q.pfam	19	0.137	15	0.138
Q.plant	52	0.099	40	0.102
NQ.pfam	19	0.132	12	0.131
NQ.plant	44	0.103	42	0.101
LG4M	13	0.140	13	0.142
LG4X	16	0.137	11	0.143

Discussion

The multi-matrix mixture models can properly handle the site rate heterogeneity among sites, therefore, outperform the single-matrix models. However, estimating multi-matrix mixture models is complicated and computational expensive. It is well-known that clade-specific models are better than general models. In this paper, we estimated both time reversible and time-none reversible 4-matrix mixture models for plant species using the distribution-free scheme to handle the rate heterogeneity among sites. Experiments showed that the new mixture models for plants outperformed the other models tested in building maximum likelihood trees on plant alignments. We recommend researchers to use the new mixture models to enhance the studies of plant evolution from protein sequences.

Declarations

Author Contributions

LSV designed the study. NHT and LSV wrote the manuscript. NHT implemented the scripts and carried out the experiments.

Data Availability

The datasets and script used in this paper are available at https://doi.org/10.6084/m9.figshare.24500212.v1

Competing Interests:

We declare that we have no conflict of interests.

References

1. Dang CC, Minh BQ, McShea H, et al (2022) nQMaker: Estimating Time Nonreversible Amino Acid Substitution Models. Syst Biol 71:1110–1123. https://doi.org/10.1093/sysbio/syac007
2. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8:275–282. https://doi.org/10.1093/bioinformatics/8.3.275
3. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109. https://doi.org/10.1093/molbev/msh112
4. Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol 29:2921–2936. https://doi.org/10.1093/molbev/mss112
5. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320. https://doi.org/10.1093/molbev/msn067
6. Le SQ, Gascuel O (2010) Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst Biol 59:277–287. https://doi.org/10.1093/sysbio/syq002
7. Le SQ, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. Philos Trans R Soc B Biol Sci 363:3965–3976. https://doi.org/10.1098/rstb.2008.0180
8. Minh BQ, Dang CC, Vinh LS, Lanfear R (2021) QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. Syst Biol 70:1046–1060. https://doi.org/10.1093/sysbio/syab010
9. Minh BQ, Schmidt HA, Chernomor O, et al (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
10. Quang LS, Gascuel O, Lartillot N (2008) Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24:2317–2323. https://doi.org/10.1093/bioinformatics/btn445
11. Ran JH, Shen TT, Wang MM, Wang XQ (2018) Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc R Soc B Biol Sci 285:. https://doi.org/10.1098/rspb.2018.1012

12. Richard Durbin SREAK and GM (2006) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. 1–371
13. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147. https://doi.org/10.1016/0025-5564(81)90043-2
14. Schwarz G (2007) Estimating the Dimension of a Model. Ann Stat 6:461–464. https://doi.org/10.1214/aos/1176344136
15. Wang HC, Li K, Susko E, Roger AJ (2008) A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol 8:. https://doi.org/10.1186/1471-2148-8-331
16. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851
17. Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401. https://doi.org/10.1093/oxfordjournals.molbev.a040082

Figures

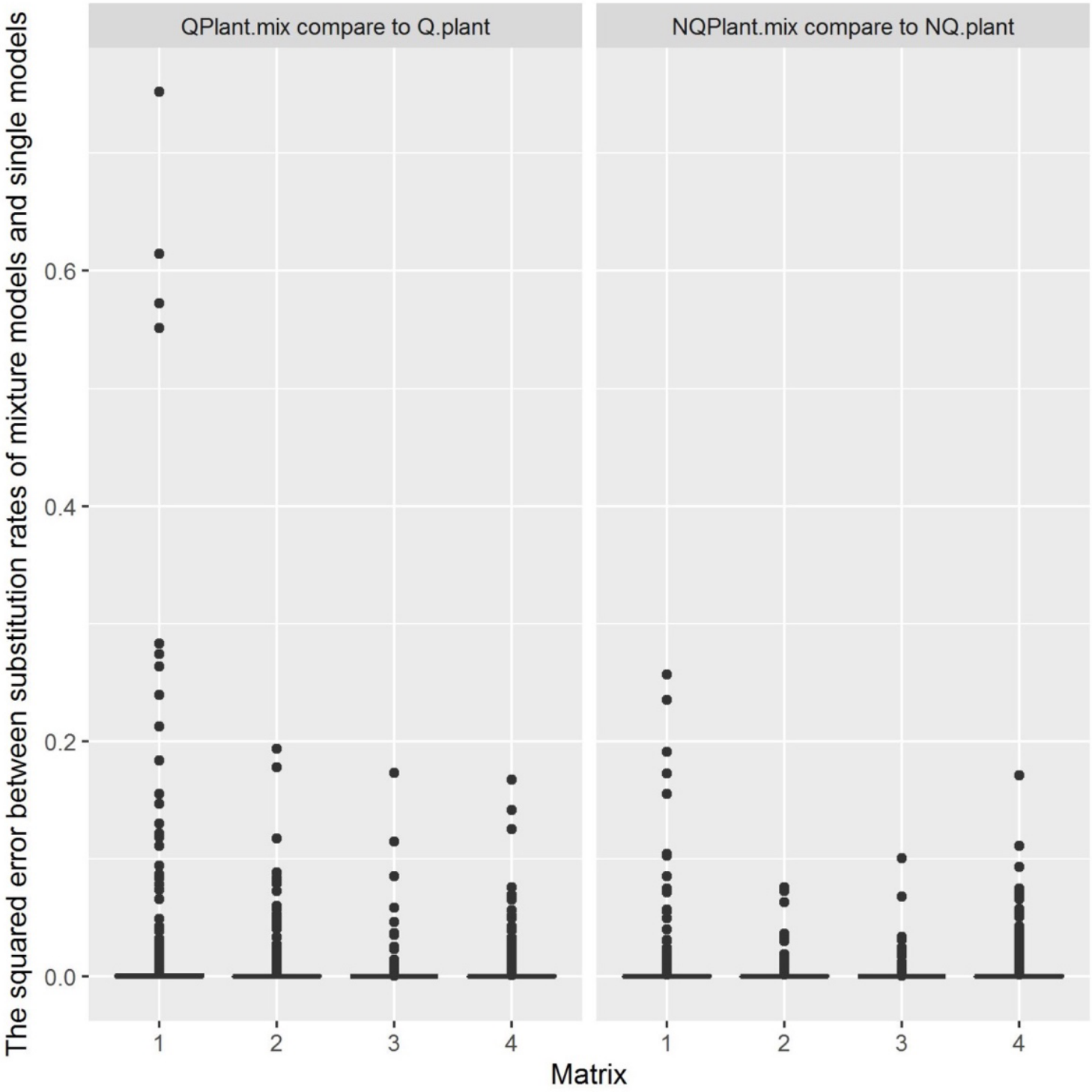

Figure 1

The square errors of substitution rates of the mixture models QPlant.mix and NQPlant.mix in comparison with the single models Q.plant and NQ.plant. Notation: 1, 2, 3, 4 are correspond to “very slow”, “slow”, “medium” and “fast” matrix of the mixture models.

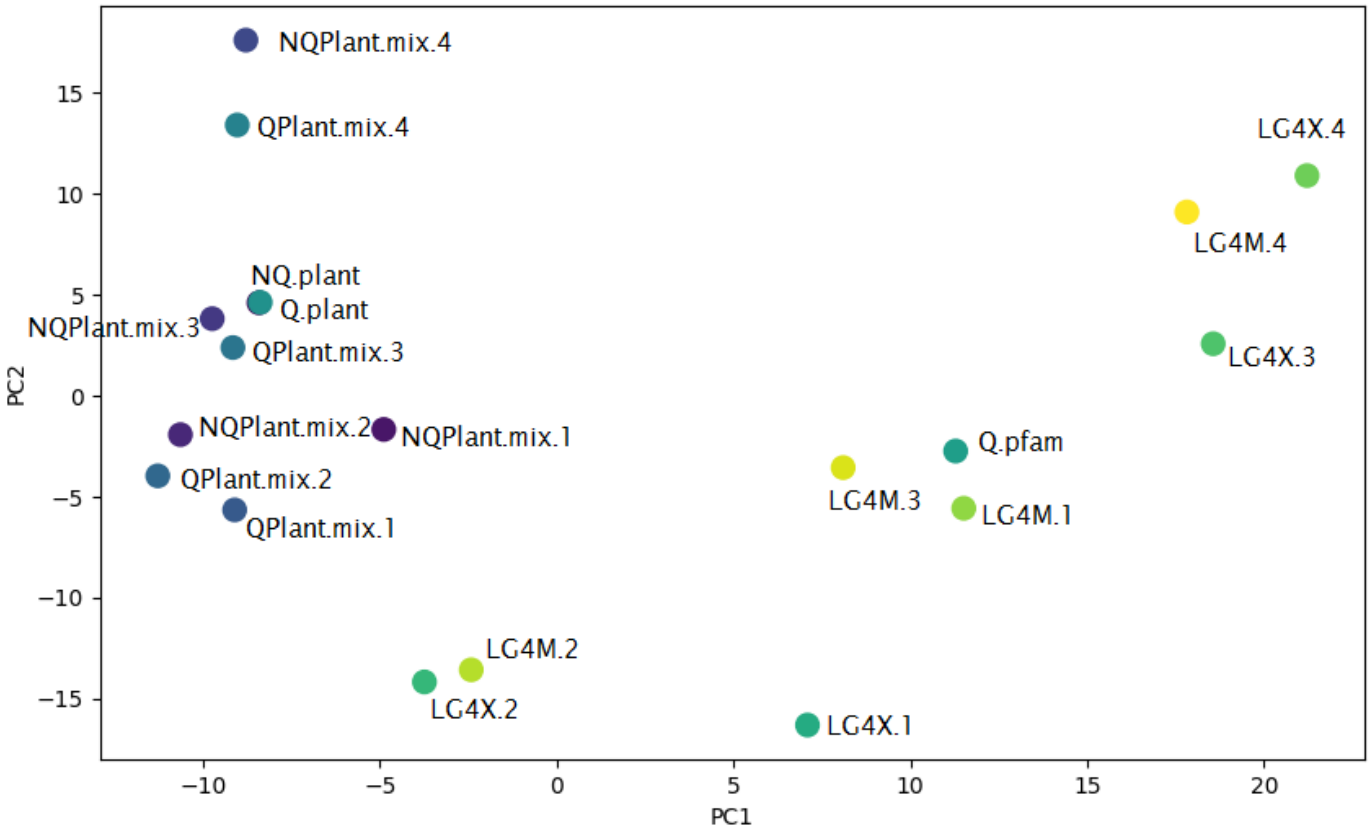

Figure 2

The principal component analysis (PCA) of models. The QPlant.mix.1(.2,.3,.4) are corresponding to “very slow” (“slow”, “medium” and “fast”) matrix. The same notations for NQPlant.mix.

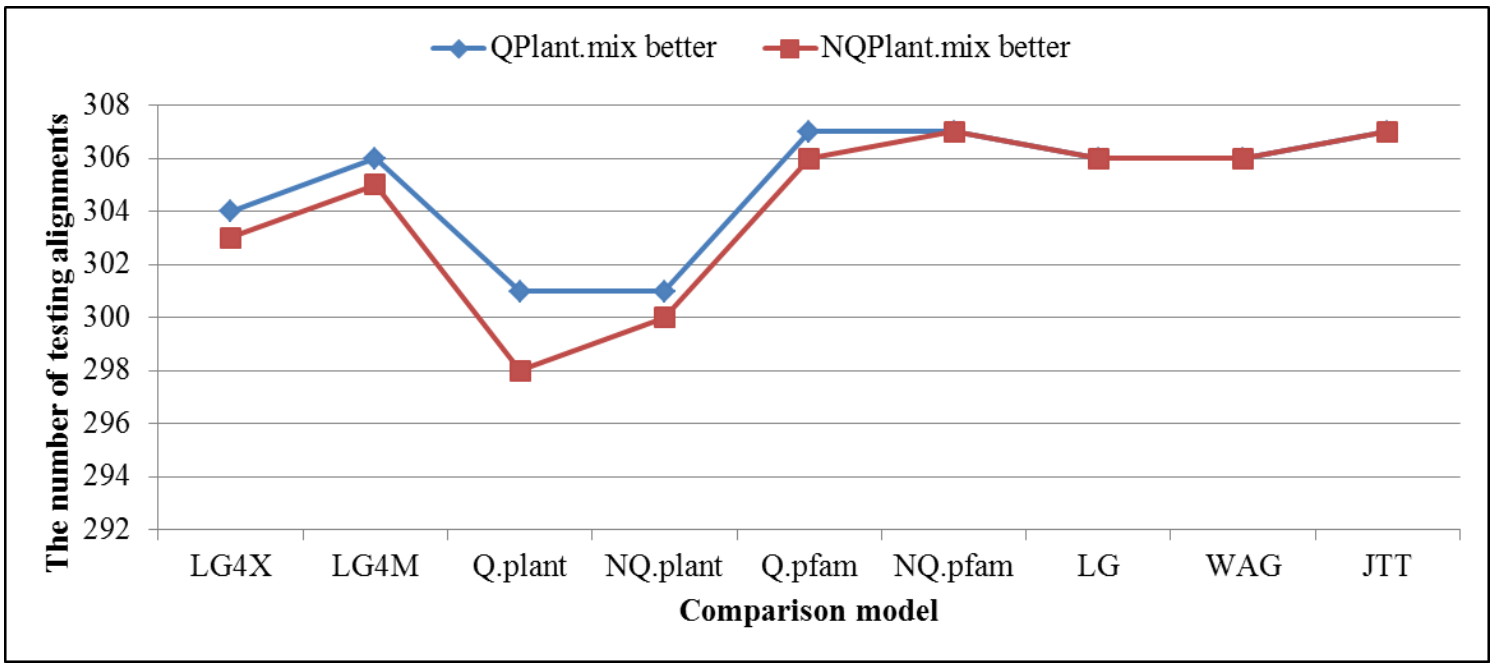

Figure 3

The performance of models in building maximum likelihood trees using the BIC criterion.