Tao et al.   (Supplementary Information)

Pervasive correlation of molecular evolutionary rates in the tree of life

Machine learning model building

Training data. We simulated nucleotide alignments using independent branch rate (IBR) and correlated branch rate (CBR) models using the NELSI package[1]. In IBR, branch-specific rates were drawn from a lognormal distribution with a mean gene rate and a standard deviation (in log-scale) that varied from 0.1 to 0.4, previously used in a study simulating independent rates with different levels of variation[1]. In CBR, branch-specific rates were simulated under an autocorrelated process[2] with an initial rate set as the mean rate derived from an empirical gene and an autocorrelated parameter, *v,* that was randomly chosen from 0.01 to 0.3, previously used in a study simulating low, moderate and high degrees of autocorrelated rates[1]. We used SeqGen[3] to generate alignments under Hasegawa-Kishino-Yano (HKY) model[4] with 4 discrete gamma categories by using a master phylogeny, consisting of 60-400 ingroup taxa randomly sampled from the bony-vertebrate clade in the Timetree of Life[5]. Mean evolutionary rates, G+C contents, transition/transversion ratios and numbers of sites for simulation were derived from empirical distributions[6]. These 2,000 simulated datasets were used as training data in building the machine learning model.

Features acquisition. Lineage-specific rate estimates (r_i's) were obtained using equations [28] - [31] and [34] - [39] in Tamura et al. (2018)[7]. As mentioned in the main text, a lineage rate is a function of all the branch rates that belong to that lineage. For any given node in the phylogeny, we extracted the relative rates of its ancestral clade (r_a) and two direct descendant clades (r_1 and r_2). Then, we calculated correlation between ancestral lineage and its direct descendant lineage rate to obtain estimates of ancestor-descendant rate correlation (ρ_{ad}). We also calculated correlation between sister lineage rates (ρ_s), for which the lineage rates of sister pairs are randomly labeled. The labeling of sister pairs have small impact on ρ_s when the number of sequences in the phylogeny is not too small (>50). However, one can also choose to resample sister pairs for multiple times and use the mean of resampled ρ_s in the CorrTest in order to eliminate any bias that may result from the arbitrary designation of sister rates during the correlation process, which can be a problem when the number of taxa is small. To avoid the assumption of linear correlation between lineages, we used Spearman rank correlation because it can capture both linear and non-linear correlation between two vectors. Two additional features derived from the relative rates in the phylogeny were used in building the machine learning model. We first estimated ρ_{ad_skip1} as the correlation between rates where the ancestor and descendant were separated by one intervening branch, and ρ_{ad_skip2} as the correlation between rates where the ancestor and descendant were separated by two intervening branches. This skipping reduces ancestor-descendant correlation, which we then used to derive the decay of correlation values by using equations $(\rho_{ad} - \rho_{ad\_skip1})/\rho_{ad}$ and $(\rho_{ad} - \rho_{ad\_skip2})/\rho_{ad}$. These two features improved the accuracy of our model slightly. In the analyses of empirical datasets, we found that a large amount of missing data (>50%) can result in unreliable estimates of branch lengths and other phylogenetic errors[8–12]. In this case, we recommend computing selected features using only those lineage pairs for which >50% of the positions contain valid data, or remove sequences with a large amount of missing data.

Predictive model. We trained a logistic regression model using the skit-learn module[13], which is a python toolbox for data mining and data analysis using machine learning algorithms, with only ρ_{ad}, only ρ_s or all 4 features ($\rho_{s,}$ $\rho_{ad,}$ the two decay of correlation features) using 2,000 simulated training datasets (1,000 with CBR model and 1,000 with IBR model). A response value of 1 was given to true positive cases (correlated rates) and 0 was assigned to true negative cases (independent rates). Thus, the prediction scores (CorrScore) were between 0 and 1. A high score representing a higher probability that the rates are correlated. Then the global thresholds at 5% and 1% significant levels can be determined. To explore the reliability of the global threshold, we re-trained the model with all 4 features extracted from 4 subsets of training data with ≤ 100 (M100), 100 – 200 (M200), 200 – 300 (M300), and > 300 (M400) sequences. A specific threshold was determined for each training subset and then was tested using Tamura *et al.* (2012)'s data[14] with the corresponding size. For example, we used the threshold determined by the model trained with small data (≤ 100 sequences) on the test data that contain less than 100 sequences, and used the threshold determined by the model trained with large data (>300 sequences) on the large test data (400 sequences). We found that the accuracy of using the specific thresholds (**Fig. S1a-c**) is similar to the accuracy when we used a global threshold (**Fig. 3d-f**). This is because the machine learning algorithm has automatically incorporated the impact of the number of sequences when it determined the relationship of four selected features (ρ_{ad}, ρ_s, and 2 decays).

Cross-validation

We performed two cross-validation tests. In 10-fold cross-validation, the predictive model was developed using 90% of the synthetic datasets, and then its performance was tested on the remaining 10% of the datasets. The AUROC was greater than 0.99 and the accuracy was high (>94%). Even in the 2-fold cross-validation, where only half of the datasets were used for training the model and the remaining half were used for testing, the AUROC was still greater than 0.99 with an accuracy greater than 92%. This indicates that the features we used in building the machine learning model are powerful and ensures high accuracy even when the training data are limited.

External tests

Publicly available data. Two previously published simulated dataset were used to evaluate CorrTest's performance. Beaulieu et al.'s data[15] contains 91 ingroup taxa with 1,000 base pairs each. For Tamura et al.'s data[14], we present the test results for the data simulated using CBR model (autocorrelated lognormal distribution) and IBR model (independent uniform distribution with 50% rate variation) here. We tested the performance of our model on CBR and IBR data with different GC contents, transition/transversion ratios, and evolutionary rates. We randomly sampled 50, 100, 200, and 300 sequences from the original 400 sequences and conducted CorrTest using the correct, error-prone topology inferred by the Neighbor-joining method[16] with an oversimplified substitution model[17]. We also tested CorrTest's performance on data simulated under an IBR model process with 100% rate variation and found that CorrTest works perfectly (100% accuracy; results not shown).

Synthetic data. We conducted another set of simulations using IBR (independent lognormal distribution) and CBR (autocorrelated lognormal distribution)[2] model with 100 replicates each using the same strategy as a training data simulation (described above) on a master phylogeny of 100 taxa randomly sampled from the bony-vertebrate clade in the Timetree of Life[5]. These 200 datasets were used to conduct CorrTest and Bayes factor analyses and to obtain the autocorrelation parameter (*v*) in MCMCTree[18].

References

1. Ho, S. Y., Duchêne, S. & Duchêne, D. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol. Ecol. Resour.* **15,** 688–696 (2015).

2. Kishino, H., Thorne, J. L. & Bruno, W. J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18,** 352–361 (2001).

3. Grassly, N. C., Adachi, J. & Rambaut, A. Seq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13,** 235–238 (1997).

4. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22,** 160–174 (1985).

5. Hedges, S. B. & Kumar, S. *The Timetree of Life*. pp3–18 (New York: Oxford University Press., 2009).

6. Rosenberg, M. S. & Kumar, S. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Mol. Biol. Evol.* **20,** 610–621 (2003).

7. Tamura, K., Tao, Q. & Kumar, S. Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates. *Mol. Biol. Evol.* msy044 (2018). doi:10.1093/molbev/msy044

8. Filipski, A., Murillo, O., Freydenzon, A., Tamura, K. & Kumar, S. Prospects for building large timetrees using molecular data with incomplete gene coverage among species. *Mol. Biol. Evol.* **31,** 2542–2550 (2014).

9. Xi, Z., Liu, L. & Davis, C. C. The impact of missing data on species tree estimation. *Mol. Biol. Evol.* **33,** 838–860 (2015).

10. Lemmon, A. R., Brown, J. M., Stanger-Hall, K. & Lemmon, E. M. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst. Biol.* **58,** 130–145 (2009).

11. Wiens, J. J. & Moen, D. S. Missing data and the accuracy of Bayesian phylogenetics. *Journal of Systematics and Evolution* **46,** 307–314 (2008).

12. Marin, J. & Hedges, S. B. Undersampling genomes has biased time and rate estimates throughout the tree of life. *Mol. Biol. Evol.* msy103 (2018). doi:doi.org/10.1093/molbev/msy103

13. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

14. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 19333–19338 (2012).

15. Beaulieu, J. M., O'Meara, B. C., Crane, P. & Donoghue, M. J. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. *Syst. Biol.* **64,** 869–878 (2015).

16. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).

17. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120 (1980).

18. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).

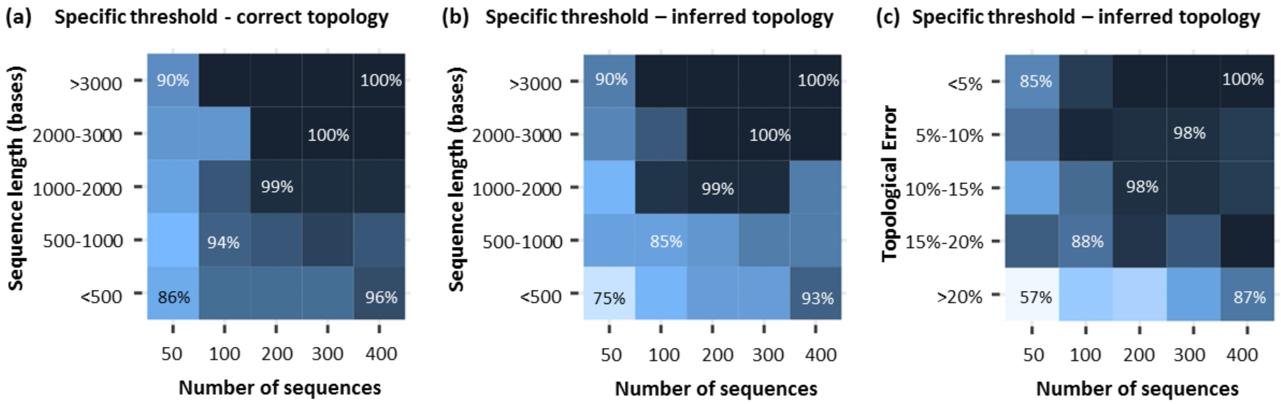

Figure S1. Patterns of CorrTest accuracy using M100, M200, M300, and M400 models for the corresponding test datasets[14]. Accuracies are shown for increasing number of sequences. The accuracy of CorrTest for different sequence length is shown when (**a**) the correct topology was assumed and (**b**) the topology was inferred. (**c**) The accuracy of CorrTest for datasets in which the inferred the topology contained small and large number of topological errors.