Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Phylogenetics

Treetime: Extending a Feature-Rich Python Library for Ancestral Sequence Reconstruction

Yinuo Jin [1] and Sanford Miller [1]

[1] Department of Computer Science, Columbia University, New York, NY 10027, USA

Abstract

Motivation: We added features relating to rate variation and site-specific matrices to TreeTime, a Python library for phylogenetic analysis, to increase the variety of sequences on which it could perform ancestral sequence reconstruction with acceptable speed and accuracy. Popular algorithms exist for joint and marginal ancestral sequence reconstruction with rate variation, but to our knowledge, they were not implemented in non-C programming library until now.

Results: With these improvements, TreeTime matched the accuracy of a state-of-the-art GUI tool for ancestral sequence reconstruction, FastML, in a comparison using a lab-grown phylogeny. The nature of a lab-grown phylogeny is that the sequences belonging to internal nodes can be known exactly, even though they are ancestral. To assess performance, we verified that TreeTime could perform both joint and marginal reconstructions of large mitochondrial sequences within a reasonable time frame (the accuracy testing only looked at marginal).

Availability: Treetime is available at https://github.com/YinuoJin/treetime

Contact: yj2589@columbia.edu and skm2159@columbia.edu

Supplementary information: Supplementary data are available at https://github.com/YinuoJin/treetime

1 Introduction

Ancestral Sequence Reconstruction (ASR) performs as a crucial step in phylogenetic analysis. Given a set of aligned sequences $S = \{s_1, ..., s_n\}$ and a tree topology τ, we can recursively reconstruct all the ancient sequences at each internal node along the phylogenetic tree. Starting from 1970s, two major categories of ASR algorithms, Maximum Parsimony (MP) & Maximum Likelihood (ML), have been proposed by multiple researchers. Nowadays ML algorithms outperform MP in reconstruction accuracy, as MP algorithms fail to handle homoplasy substitution through evolution very well [Liberles, 2007].

Multiple ASR inference softwares were developed and distributed in the past decades, including FASTML [Ashkenazy et al., 2012], PhyML [Oliva et al., 2019], PAML [Yang, 1997], etc. Nevertheless, all of them are online servers or command-line toolkits written in C, which are either inconvenient for non-primary developers to update features or not interactive to users. Recently, a comprehensive python library for phylogenetic analysis called TreeTime was developed by Sagulenko et al. [2018]. ASR in Treetime serves as an important step for downstream analysis, but the current implementation only supports the most vanilla algorithm. Therefore we decided to extend ASR features to incorporate with the up-to-date inference algorithms.

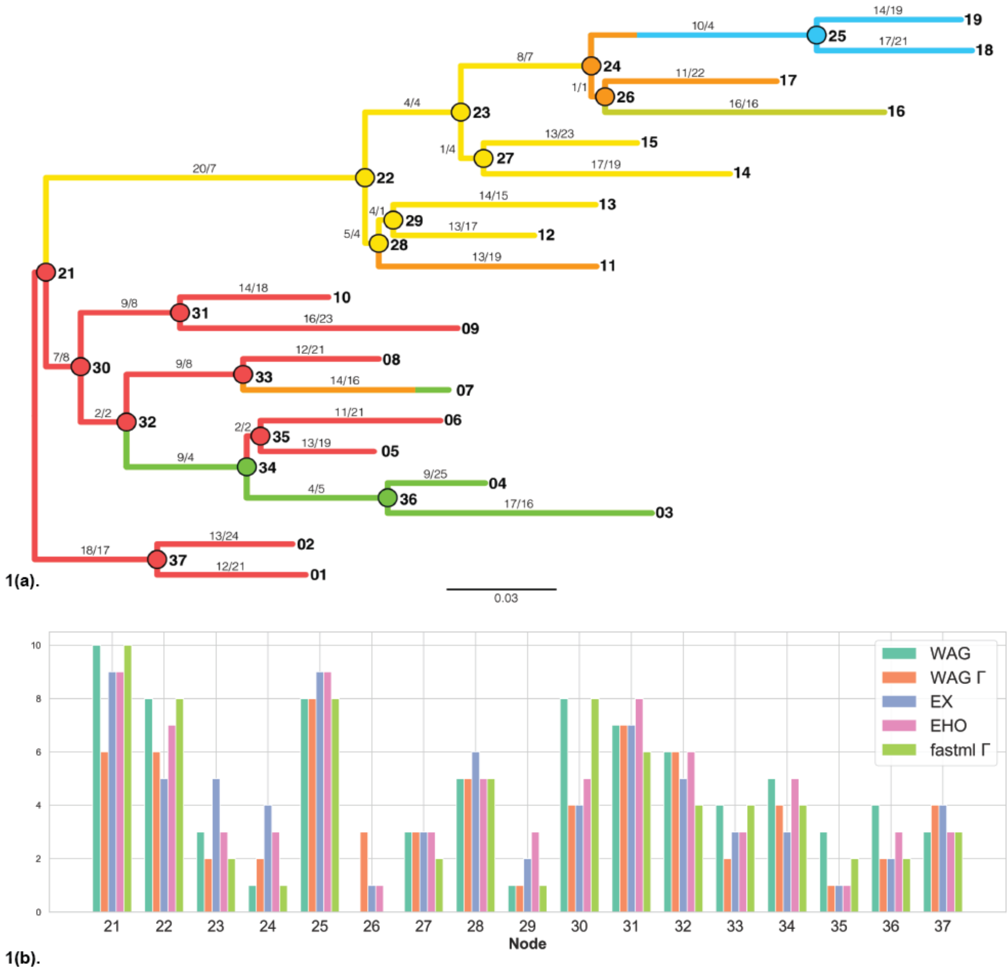
2 Methods

2.1 Marginal Reconstruction

Marginal Reconstruction calculates the most likely sequences at a certain node given the characters at all other nodes. Treetime [Sagulenko et al., 2018] has implemented the optimal algorithm utilizing dynamic programming, which can recover the sequences at all ancestral nodes in linear time with number of extant sequences [Felsenstein, 1981]. However, such algorithm assumes fix substitution rate and independent evolution at different sites during the reconstruction, and thus couldn't reflect the heterogenous mutation nature and functional dependencies along the sequence [Pupko et al., 2002]. Moshe and Pupko [2019] and Ishikawa et al. [2019] proposed the efficient marginal reconstructions with among-site rate variations (ASRV) almost at the same time, and we followed their guideline to extend the vanilla marginal reconstructions in Treetime.

2.1.1 Among-site Rate Variation (ASRV)

Previous researchers [Yang, 1996, Pupko et al., 2002] have found that a single, fixed substitution rate for ASR can't fit the sequence data very well. In particular, evolutionary and functional constraints cause the non-homogeneous mutations at different positions of sequences, such as different codon positions in DNA coding regions [Yang, 1996] and stem/loop structure in tRNA sequences. Yang [1994] suggested using a single-parameter gamma distribution to model the mutation heterogeneity



1(a).

1(b).

Fig. 1. Number of incorrectly inferred sites for ASR on FP dataset ($\alpha = 0.01$
1(a). The original ground tree topology provided by [Randall et al., 2016] 1(b). Number of incorrectly inferred sites on FP dataset at each ancestral node; Methods for comparison include WAG (vanilla model with the fixed rate), WAG + Γ (ASRV), EX (ASRV + structural information with solvent accessibility), EHO (ASRV + structural information with protein secondary profile), FASTML (results from FASTML, the standard state-of-art ASR server

at different sites [1].

Assume the observable extant sequences E and tree topology τ, to reconstruct the most likely character at position c at any arbitrary node v is to calculate $\text{argmax}_a Pr(V_c = a \mid E, \tau)$. The marginal reconstruction is obtained with Bayes theorem:

$$Pr(V_c = a \mid E, \tau) = \frac{Pr(E \mid V_c = a, \tau)Pr(V_i = a)}{Pr(E)} \tag{1}$$

Reconstructions at all ancestral sites V_i in Equation (1) could be solved by the efficient dynamic programming algorithm proposed by Moshe and Pupko [2019] and Ishikawa et al. [2019] with a two-step tree traversals (bottom-up propagation & top-down character assignments) [2]. To incorporate with ASRV, we applied the discrete-gamma model (Yang, 1996) with 8 categories to approximate the gamma distribution. Let $R = \{r_1, .., r_8\}$ with $Pr(r_i) = \frac{1}{8}$:

$$Pr(V_c = a \mid E, \tau, R) = \frac{\sum_i Pr(E \mid V_c = a, \tau, r_i)Pr(r_i)Pr(V_c = a)}{Pr(E)} \tag{2}$$

Utilizing the DP algorithm averaging likelihood across all candidate rates (Equation (2)), we implemented the marginal reconstruction with ASRV by imitating Moshe and Pupko [2019]'s algorithm.

2.1.2 Structural Constraints

Aside from ASRV to model the sequence evolvement heterogeneity, structural constraints is an extra information for ASR. Koshi and Goldstein [1995] created a model structural information of protein sequences for ASR inference. However, it wasn't widely used due to limited protein 3D information. More recently, Le and Gascuel [2010] proposed new site-specific models with more comprehensive training set. They introduced two separate mixture model, EX and EHO to capture the 2-state in solvent accessibility ('Exposed' or 'Buried'), or 3-state in secondary structure

[1] when $\alpha <= 1$, the L-shaped curve indicates that most sites are invariant and fewer sites are highly mutable; when $\alpha > 1$, the bell-shaped curve indicates that more sites have similar mutation rates

[2] For detailed algorithm summary, please refer to Pupko *et al.* (2019) or supplementary documents in https://github.com/YinuoJin/treetime

('Extended', 'Helix' & 'Coil') respectively [3].

Moshe and Pupko [2019] incorporated with the site-specific mixture as another further step of ASRV. We extended this feature in Treetime as well. In general, the extra information provided is a set of substitution matrices $M_1, ..., M_k$ with a string of properties $p_1, ..., p_n$indicating the classified state at each site of the sequence (e.g. 'EBBBEBEEB' or 'EHCHHHCEE'). In general, the inference is to average over all candidate rate r_i and substitution matrix M_j:

$$Pr(V_c = a \mid E, \tau, R, M) = \frac{\sum_i \sum_j Pr(E \mid V_c = a, \tau, r_i, M_j) Pr(r_i) Pr(M_j) Pr(V_c = a)}{Pr(E)} \quad (3)$$

Following the method of Moshe and Pupko [2019], equation (3) can still be solved in linear time with respect to number of extant nodes, as ASRV provides constant ($|R| = 8$) number of extra calculation at each node, and mixture models require an extra constant ($|M| = 2$ or 3). Similar to $Pr(r_i) = \frac{1}{8}$ in ASRV marginal inference, $Pr(M_j)$ is assigned as either 1 or 0 depending on the property of the site. The total runtime is estimated as $T(n) = 2 \cdot |R| \cdot |M| \cdot n <= 48n \sim O(n)$. The multiplier "2" indicates the two-step tree traversal, where each node is visited twice during the marginal ASR.

2.2 Joint Reconstruction

Joint reconstruction calculates the set of sequences for the internal nodes of a phylogenetic tree that produce the most likely tree overall. The differing aims of joint and marginal reconstruction cause them to produce somewhat different sequences [Pupko et al., 2000]. Naive attempts to account for rate variation in the calculation of joint reconstruction lead to exponential complexity, since, unlike in marginal reconstruction, each node's optimal choice of value depends on the choices of value made at every other node–which entails that an unsophisticated approach would have to calculate every possible combination of nodes and values. Pupko et al. [2002] propose a branch-and-bound algorithm to circumvent this issue. They develop a technique to identify an upper bound on the likelihood of all members of a group of potential trees; if any previously-found reconstruction has a higher likelihood than the maximum possible likelihood calculated for the group, that swath of the search space can be ruled out *en masse*. This approach is speed up by searching the most likely marginal reconstruction first, which are likely to have high likelihoods, so large numbers of alternatives can be quickly eliminated afterwards. Branch-and-bound obtains provably optimal reconstructions. While technically exponential in the worst case, it takes only approximately linear time in practice. The TreeTime implementation of branch-and-bound has no significant modifications from the original algorithm.

3 Results

We tested our implementation step wise with unittest with simulations and real dataset. The unittest with simulation sequences are included in the "test" directory in our github page. We mainly check the correctness of our implementations with simulations. For joint reconstructions, we ran a simple runtime benchmark due to the exponential complexity of the algorithm (Figure 2). For marginal reconstructions, we evaluated our implementation with real datasets: fluorescent protein (FP) sequences and mitochondrial protein sequences.

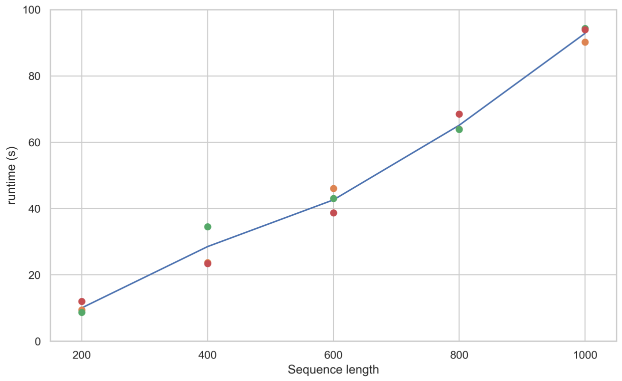


Fig. 2. Runtime benchmarking with joint reconstruction implementations with ASRV. A simple topology with 4 sequences is fixed with varying length of sequence lengths. Time increases linearly as length of the sequence increases. Under a shallow tree topology, the joint ASRV has acceptable running time withing a few minutes

Table 1. Number of incorrectly inferred ancestral sites for FP datase

α	WAG (fixed)	WAG + Γ	EX	EHO	FastML[a]
0.01	79	68	74	79	71
0.545	79	79	75	82	71

[a] The value FastML = 71 is extracted from the paper Randall et al. [2016]

3.1 Fluorescent Protein Experimental phylogeny

The FP sequences are xperimentally generated phylogeny from Randall et al. [2016]. It's the only dataset we could find with ground truth ancestral sequences and tree topology. We evaluated our results with the original vanilla model (WAG), our ASRV model (WAG + Γ) and two site-specific mixture models. We obtained the α parameter from two sources: (1). estimate from RaxML Stamatakis, 2014 during the tree inference. (2). Parameter fine-tuning on α (0.01 - 1.5) against ground truth. We selected the highest α value (0.01). For site-specific properties of the sequences, we inferred the solvent accessibility and secondary structure from the primary protein sequences with Sable [Adamczak et al., 2004], a neural-networks based software for protein structural information inference.

We compared our results against the vanilla marginal reconstruction (WAG (fixed)) and the state-of-art FastML server. ASRV and mixture model with solvent accessibility outperform the original model with fixed rate, and their results under either α value selection (Table 1 & Figure 1) are close with the FastML. The mixture model EHO using protein secondary structure obtained the worse result, which might be caused by the errors during secondary structure prediction. Ideally, if we confidently know the 3D structure of the sequences, it should have similar accuracy with EX. Figure 1 shows the detailed number of incorrectly reconstructed sites at internal nodes, where the errors aren't homogeneous under different models.

Our results aren't identical to FastML, which could be caused by the following reasons: (1). FastML estimates α using expectation maximization, and Randall et al. [2016] didn't cite their estimated α value. (2). Treetime instantiates a GTR model with μ (the average mutation rate), and we used the default value $\mu = 1$, which might not be optimal to the dataset.

[3] In either mixture model, distinguished substitution matrices are associated with each category: e.g. 'B' matrix and 'E' matrix are given for the 'EX' model

3.2 Mitochondrial datasets

As mentioned by Moshe and Pupko [2019], the experimental FP dataset is a simple and highly homogenous phylogeny with only 19 extant nodes with $l = 225$. To further assess the consistency of our methods, we compared the likelihood of different models on two separate sets of primate mitochondrial protein sequences (Table 2). All the sequences were collected from NCBI protein database.

Among the dataset, 80 cytochrome oxidase subunit I sequences and 52 NADH dehydrogenase subunit 5 sequences (both are protein sequences) were randomly collected out of distinct primate species. Similar to the FP dataset analysis, experiments were performed with WAG (original Treetime implementation), WAG + Γ, EX and EHO models. We performed MSA with Pasta [Mirarab et al., 2015], and inferred α parameter and topology with RAxML [Stamatakis, 2014]. Performances were evaluated with loglikelihood scores at root nodes. In both datasets, ASRV and mixture models with structural constraints yielded higher likelihood scores than the plain fixed rate model (Table 3, Table 4)

The pure ASRV has slightly higher likelihood scores than ASRV + mixture models, and among mixture models EX always has higer score than EHO, which is consistent with of [Moshe and Pupko, 2019]'s results with their own dataset. Figure 3 presents the distributions of likelihood at each site with different models, where the figures in the right panels show the loglikelihood score differences between each model and a baseline LG (fixed) model [5].

Table 2. Dataset information in the current experiment

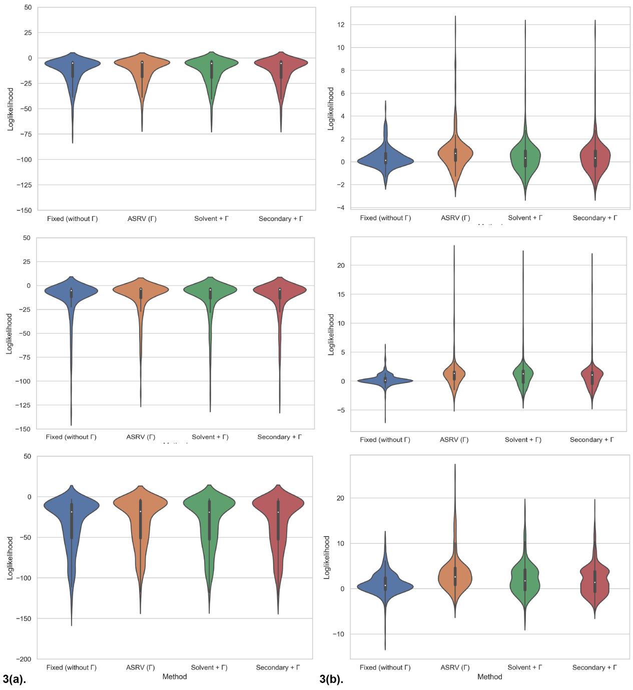
Dataset	Seq-length	α	Branch length
cytochrome oxidase subunit I	527	0.248	1.714
NADH dehydrogenase subunit 5	604	0.487	5.060

Table 3. Total root loglikelihood scores for each dataset

Dataset	WAG (fixed)	WAG + Γ	EX	EHO
cytochrome oxidase subunit I	-7339.51	-7204.33	-6576.55	-6884.19
NADH dehydrogenase subunit 5	-19889.90	-18523.83	-19234.11	-19439.74

Table 4. Percent of root sites with better likelihood scores than WAG (fixed)

Dataset	WAG + Γ	EX	EHO
cytochrome oxidase subunit I	77.04%	66.03%	58.63%
NADH dehydrogenase subunit 5	72.85%	61.75%	57.12%



3(a). 3(b).

Fig. 3. Distributions of loglikelihood scores of each model (fixed, ASRV, ASRV+EX, ASRV+EHO) on three datasets; the upper panel for FP dataset, the lower 2 panels for mitochondrial dataset. Figure 3(a). (the left panel) shows the original likelihood scores, 3(b). shows the comparison against a baseline LG model. Both ASRV and ASRV+mixture models show higher likelihood with respect to fixed model

Although loglikelihood scores can't always reflect the accuracy, the consistent performances from 3 independent datasets show that incorporating with ASRV and site-specific constraints in protein ASR might be a good approach for inferring more accuracy reconstructions for future research.

4 Discussion

We extended multiple new features / models with the idea of among-site rate variations (ASRV) and site-specific constraints (via mixture models) for ancestral sequence reconstructions to Treetime, a novel python library for phylogenetic analysis. Through tests with simulation data and evaluations with real data on marginal reconstructions, we found that ASRV and mixture models of protein sequences can improve the ASR accuracy if α and structural information could be obtain / inferred properly. Nevertheless, inprecise inference of structural information, especially secondary structures may worsen the accuracy as well. We further built mixture models upon branch & bound joint ASRV, which is a novel feature from known softwares. Nevertheless, due to the constraints of exponential runtime, we haven't yet to evaluate the feasibility of this feature.

To further generalize the the idea of functional / structural dependencies on ASR, further information such as DNA assembly, tRNA secondary structure Rzhetsky [1995] and codon models may be explored as well (communication with Prof. Pe'er); more complicated, experimental benchmark dataset in the future should be used to testify the robostness of mixture models [Moshe and Pupko, 2019]. In recent years, researchers have proposed Minimum Posterior Estimated Error (MPEE) aside of Maximum Likelihood [Oliva et al., 2019] as the metrics to evalute performance, and Gamma mixture models [Mayrose et al., 2005] to

[5] This comparison methods was suggested by [Moshe and Pupko, 2019] for clearer view of 'superiority' percentage of each model against a baseline.

model substitution stochasticity for ASR. Treetime, as a scalable, well-documented python library has the potential to explore those new methods as well.

References

Rafał Adamczak, Aleksey Porollo, and Jarosław Meller. Accurate prediction of solvent accessibility using neural networks–based regression. *Proteins: Structure, Function, and Bioinformatics*, 56(4):753–767, 2004.

Haim Ashkenazy, Osnat Penn, Adi Doron-Faigenboim, Ofir Cohen, Gina Cannarozzi, Oren Zomer, and Tal Pupko. Fastml: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic acids research*, 40(W1):W580–W584, 2012.

Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

Sohta A Ishikawa, Anna Zhukova, Wataru Iwasaki, and Olivier Gascuel. A fast likelihood method to reconstruct and visualize ancestral scenarios. *Molecular biology and evolution*, 36(9):2069–2085, 2019.

Jeffrey M Koshi and Richard A Goldstein. Context-dependent optimal substitution matrices. *Protein Engineering, Design and Selection*, 8(7):641–645, 1995.

Si Quang Le and Olivier Gascuel. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Systematic Biology*, 59(3): 277–287, 2010.

David A Liberles. *Ancestral sequence reconstruction*. Oxford University Press on Demand, 2007.

Itay Mayrose, Nir Friedman, and Tal Pupko. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, 21(suppl_2):ii151–ii158, 2005.

Siavash Mirarab, Nam Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. Pasta: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, 22(5):377–386, 2015.

Asher Moshe and Tal Pupko. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. *Bioinformatics*, 35 (15):2562–2568, 2019.

Adrien Oliva, Sylvain Pulicani, Vincent Lefort, Laurent Brehelin, Olivier Gascuel, and Stéphane Guindon. Accounting for ambiguity in ancestral sequence reconstruction. *Bioinformatics*, 35(21):4290–4297, 2019.

Tal Pupko, Itsik Pe, Ron Shamir, and Dan Graur. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular biology and evolution*, 17(6):890–896, 2000.

Tal Pupko, Itsik Pe'er, Masami Hasegawa, Dan Graur, and Nir Friedman. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics*, 18(8):1116–1123, 2002.

Ryan N Randall, Caelan E Radford, Kelsey A Roof, Divya K Natarajan, and Eric A Gaucher. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nature communications*, 7(1):1–6, 2016.

Andrey Rzhetsky. Estimating substitution rates in ribosomal rna genes. *Genetics*, 141(2):771–783, 1995.

Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018.

Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

Ziheng Yang. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3):306–314, 1994.

Ziheng Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.

Ziheng Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556, 1997.