MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods

Koichiro Tamura,[1,2] Daniel Peterson,[2] Nicholas Peterson,[2] Glen Stecher,[2] Masatoshi Nei,[3] and Sudhir Kumar*[,2,4]

[1]Department of Biological Sciences, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

[2]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University

[3]Department of Biology and the Institute of Molecular Evolutionary Genetics, The Pennsylvania State University

[4]School of Life Sciences, Arizona State University

***Corresponding author:** E-mail: s.kumar@asu.edu.

Associate editor: Naoki Takebayashi

Abstract

Comparative analysis of molecular sequence data is essential for reconstructing the evolutionary histories of species and inferring the nature and extent of selective forces shaping the evolution of genes and species. Here, we announce the release of Molecular Evolutionary Genetics Analysis version 5 (MEGA5), which is a user-friendly software for mining online databases, building sequence alignments and phylogenetic trees, and using methods of evolutionary bioinformatics in basic biology, biomedicine, and evolution. The newest addition in MEGA5 is a collection of maximum likelihood (ML) analyses for inferring evolutionary trees, selecting best-fit substitution models (nucleotide or amino acid), inferring ancestral states and sequences (along with probabilities), and estimating evolutionary rates site-by-site. In computer simulation analyses, ML tree inference algorithms in MEGA5 compared favorably with other software packages in terms of computational efficiency and the accuracy of the estimates of phylogenetic trees, substitution parameters, and rate variation among sites. The MEGA user interface has now been enhanced to be activity driven to make it easier for the use of both beginners and experienced scientists. This version of MEGA is intended for the Windows platform, and it has been configured for effective use on Mac OS X and Linux desktops. It is available free of charge from http://www.megasoftware.net.

Introduction

The Molecular Evolutionary Genetics Analysis (MEGA) software was developed with the goal of providing a biologist centric, integrated suite of tools for statistical analyses of DNA and protein sequence data from an evolutionary standpoint. Over the years, it has grown to include tools for sequence alignment, phylogenetic tree reconstruction and visualization, testing an array of evolutionary hypotheses, estimating sequence divergences, web-based acquisition of sequence data, and expert systems to generate natural language descriptions of the analysis methods and data chosen by the user (Kumar et al. 1994, 2008; Kumar and Dudley 2007). With the fifth major release, the collection of analysis tools in MEGA has now broadened to include the maximum likelihood (ML) methods for molecular evolutionary analysis. Table 1 contains a summary of all statistical methods and models in MEGA5, with new features marked with an asterisk (*). In the following, we provide a brief description of methodological advancements, along with relevant research results, and technical enhancements in MEGA5.

Model Selection for Nucleotide and Amino Acid sequences

MEGA5 now contains facilities to evaluate the fit of major models of nucleotide and amino acid substitutions, which are frequently desired by researchers (Posada and Crandall 1998; Nei and Kumar 2000; Yang 2006) (fig. 1A). For nucleotide substitutions, the GTR and five nested models are available, whereas six models with and without empirical frequencies (+F) have been programmed for the amino acid substitutions (Table 1). MEGA5 provides the goodness-of-fit (see below) of the substitution models with and without assuming the existence of evolutionary rate variation among sites, which is modeled by a discrete Gamma distribution (+G) (Yang 1994) and/or an allowance for the presence of invariant sites (+I) (Fitch and Margoliash 1967; Fitch 1986; Shoemaker and Fitch 1989). This results in an evaluation of 24 and 48 models for nucleotide and amino acid substitutions, respectively. For each of these models, MEGA5 provides the estimated values of shape parameter of the Gamma distribution (α), the proportion of invariant sites, and the substitution rates between bases or residues, as applicable. Depending on the model, the assumed or observed values of the base or amino acid frequencies used in the analysis are also provided. This information enables researchers to quickly examine the robustness of the estimates of evolutionary parameters under different models of substitutions and assumptions about the distribution of evolutionary rates among sites (fig. 1C). The goodness-of-fit of each model to the data is measured by the Bayesian information criterion (BIC, Schwarz 1978) and corrected

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Research resource

http://mbe.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

Table 1. A Summary of Analyses and Substitution Models in MEGA5

Sequence alignments
DNA, codon, and protein alignments; both manual and automated alignments with trace file Editor. Built-in automated aligners: CLUSTALW and MUSCLE*.
Major analyses (statistical approach in parentheses)
Models and parameters: Select Best-Fit Substitution Model* (ML); test pattern homogeneity; Estimate Substitution Pattern (MCL, ML*); Estimate Rate Variation Among Sites* (ML); Estimate Transition/Transversion Bias (MCL, ML*); Estimate Site-by-Site Rates* (ML).
Infer phylogenies: Infer Phylogenetic Trees (NJ, ML*, ME, MP); Phylogeny Tests (Bootstrap and Branch-length tests); Branch-and-Bound Exact Search (MP); Heuristic Searches: Nearest-Neighbor-Interchange (NNI; ML*, ME, MP), Close-Neighbor-Interchange (CNI; ML*, ME, MP), and Max–Mini (MP)
Compute distances: Pairwise and Diversity; Within- and Between-Group Distances; Bootstrap and Analytical Variances; separate distances by Site Degeneracy, Codon Sites; Separation of Distances in Transitions and Transversions; Separate Nonsynonymous and Synonymous Changes
Tests of Selection: For Complete Sequences or Set of Codons; Sequence Pairs or Groups (Within and Between)
Ancestral Sequences: Infer by ML with Relative Probabilities for bases or residues* or by MP (all parsimonious pathways)
Molecular Clocks: Tajima's 3-Sequence Clock Test*; Likelihood Ratio Test (ML) for a Topology*; Estimate Branch Lengths under Clock*
Substitution models (+F = with empirical frequencies; REV = reversible)
DNA: General Time Reversible (GTR)*, Tamura–Nei, Hasegawa–Kishino–Yano*, Tamura Three-Parameter, Kimura Two-Parameter, Tajima–Nei, Jukes–Cantor
Codons: Nei–Gojobori (original and modified), Li–Wu–Lou (original and modified)
Protein: Poisson, Equal-Input, Dayhoff (+F), Jones–Taylor–Thornton (+F), Whelan and Goldman (+F)*, Mitochondrial REV (+F)*, Chloroplast REV (+F)*, Reverse Transcriptase REV (+F)*
Rate Variation and Base Compositions: Gamma rates (G) and Invariant sites (I)* models; Incorporate Compositional Heterogeneity.

NOTE.—MCL, maximum composite likelihood; ME, minimum evolution; MP, maximum parsimony. An asterix (*) denotes features that are new in MEGA5.

Akaike information criterion (AICc, Hurvich and Tsai 1989) (see also Posada and Buckley 2004). By default, MEGA5 lists models with decreasing BIC values (see below for the reason and caveats), along with log likelihood as well as AICc values for each model.

In the ML methods for evaluating the fit of substitution models to the data, an evolutionary tree is needed. MEGA5 automatically infers the evolutionary tree by the Neighbor-Joining (NJ) algorithm that uses a matrix of pairwise distances estimated under the Jones–Thornton–Taylor (JTT) model for amino acid sequences or the Tamura and Nei (1993) model for nucleotide sequences (Saitou and Nei 1987; Jones et al. 1992; Tamura and Nei 1993; Tamura et al. 2004). Branch lengths and substitution rate parameters are then optimized for each model to fit the data. Users may provide their own tree topology in the Newick (New Hampshire) format for use in this model selection (fig. 1*B*). However, the automatic option is expected to be frequently used because trees are rarely known *a priori*. We tested the impact of the use of automatically generated trees in MEGA5 on the process of model selection by computer simulation. These simulations used 448 sets of evolutionary parameters (base frequencies, sequence length, mean evolutionary rate, and transition–transversion rate ratio) derived from real sequence data (see Rosenberg and Kumar 2001) and introduced four different levels of rate variation among sites for each parameter set (Gamma shape parameter, α = 0.25, 0.5, 1.0, and 2.0). Results showed that the best-fit models produced by using automatically generated trees were the same as those inferred using the true tree for $\geq$93% of the data sets according to the BIC and AICc criteria (fig. 2*A*).

For an overwhelming majority of data sets, AICc selected the most complex model (see also Ripplinger and Sullivan 2008). But, both BIC and AICc selected substitutions models that were more complex than the true model (Posada and Buckley 2004; Alfaro and Huelsenbeck 2006). The true model was among the top-3 when BIC was used and among the top-5 when AICc was used. When the rate variation among sites was extreme (α = 0.25), models incorporating invariant sites (+I) along with discrete gamma rate categories (+G) were favored for virtually every data set. This means that a discrete gamma (+G) model using a small number of categories (4), which is a common practice, coupled with an allowance for invariant sites (+I) is better at approximating the continuous Gamma distribution used in the simulation when the rate variation among sites is severe. This was confirmed by comparing the ML value for the fit of HKY + G model (10 categories) with the ML value for GTR + G + I model using only four discrete gamma categories. The former performed slightly better than the latter, even though the latter involved a more complex model.

On the basis of the above observation, we pooled +G and +G + I results for each model of substitution and found that BIC selects the true model for >70% of the data sets. In contrast, AICc selects the correct model only 35% of the time. Therefore, we rank the models by BIC in MEGA5 (fig. 1*C*). However, the choice of criterion to select the best-fit models is rather complicated, and researchers should explore model selection based on AICc values and other available methods for evolutionary analyses in which choice of model is known to substantially affect the final result (e.g., Tamura et al. 2004; Ripplinger and Sullivan 2010). To facilitate downstream analysis to select the best model, MEGA5 provides exporting of results in Microsoft Excel/Open Office and comma separated values formats.

These simulation results also provided us with an opportunity to evaluate the estimates of α obtained by using the automatically generated tree and to compare them to those obtained by using the true tree under the correct model of substitution. The means and standard deviations of these estimates were very similar to the true values and virtually identical for automatically generated and true trees (fig. 2*B*). Similarly, the overall estimates of

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

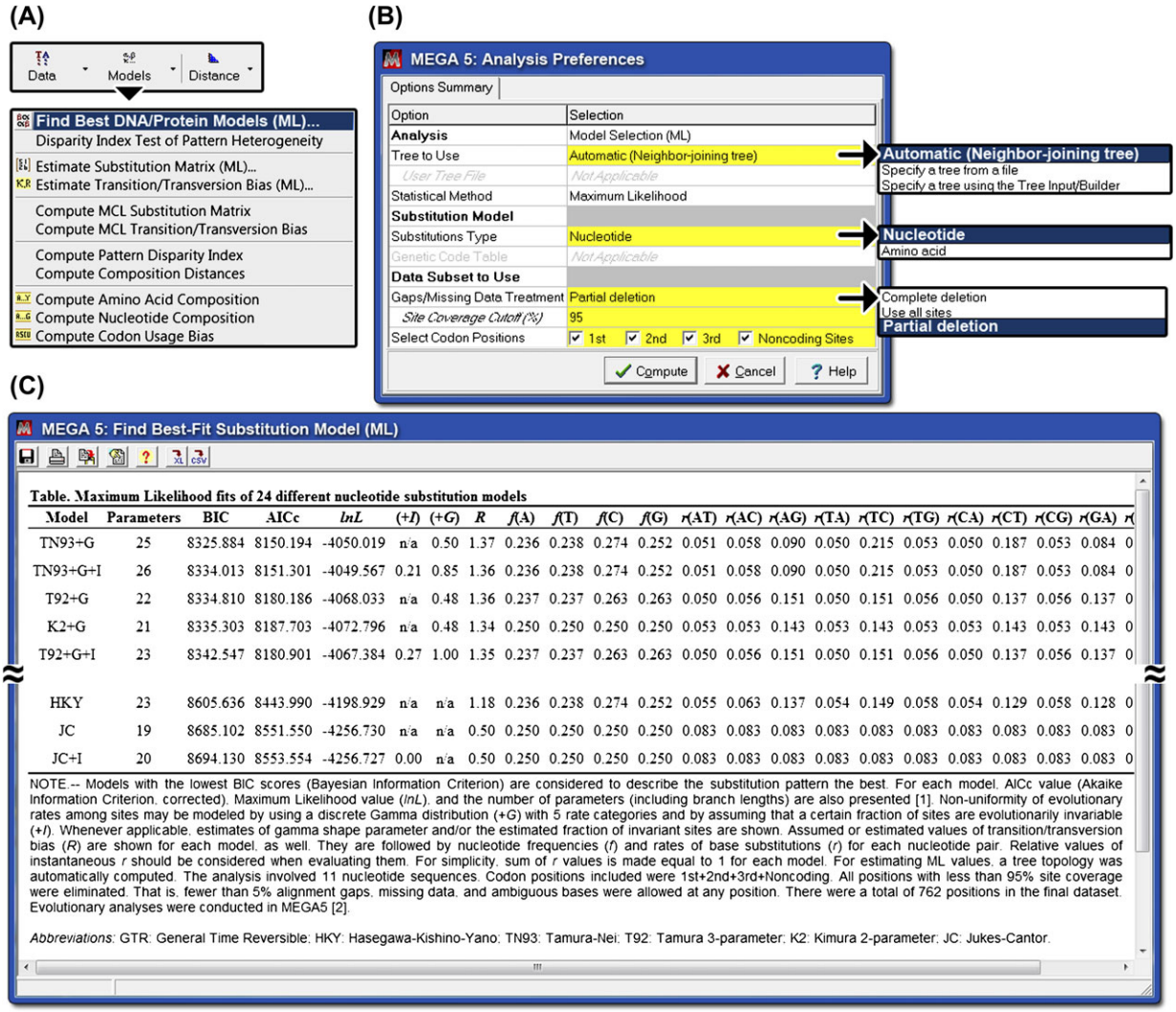

FIG. 1. Evaluating the fit of substitution models in MEGA5. (*A*) The "Models" menu in the "Action Bar" provides access to the facility. (B) An "Analysis Preferences" dialog box provides the user with an array of choices, including the choice of tree to use and the method to treat missing data and alignment gaps. In addition to the "Complete Deletion" and "Pairwise Deletion" options, MEGA5 now includes a "Partial Deletion" option that enables users to exclude positions if they have less than a desired percentage (x%) of site coverage, that is, no more than (100−x)% sequences at a site are allowed to have an alignment gap, missing datum, or ambiguous base/amino acid. For protein coding nucleotide sequences, users can choose to analyze nucleotide or translated amino acid substitutions, with a choice of codon positions in the former. (*C*) The list of evaluated substitution models along with their relative fits, number of parameters (branch lengths + model parameters), and estimates of evolutionary parameters for *Drosophila* Adh sequence data which are available in the Examples directory in MEGA5 installation. The note below the table provides a brief description of the results (e.g., ranking of models by BIC), data subset selected, and the analysis option chosen. This figure is available in color online and in black and white in print.

transition–transversion ratio (R) were close to the true value for both automatically generated and true trees (fig. 2C). Therefore, the use of automatically generated trees with MEGA5 is useful, as a first approximation, in estimating evolutionary substitution parameters and evaluating relative fits of models.

Inferring ML Trees

MEGA5 now provides the ML method to infer evolutionary trees and conduct the bootstrap test for nucleotide and amino acid alignments (Felsenstein 1981, 1985). Because the ML method is computationally demanding, we provide heuristic methods that search for the ML tree by topological rearrangements of an initial tree (Swofford 1998; Nei and Kumar 2000; Guindon and Gascuel 2003; Stamatakis et al. 2005). The initial tree for the ML search can be supplied by the user (Newick format) or generated automatically by applying NJ and BIONJ algorithms to a matrix of pairwise distances estimated using a maximum composite likelihood approach for nucleotide sequences and a JTT model for amino acid sequences (Saitou and Nei 1987; Jones et al. 1992; Gascuel 1997; Tamura et al. 2004). For the user-selected data subset that contains sites with insertion–deletions and missing data, we begin by temporarily obtaining a site coverage parameter such that the number of ambiguous states and insertion–deletions per sequence are the lowest. This site coverage parameter is then used to generate a data subset for estimating evolutionary distances to build an initial tree along with branch lengths. We

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

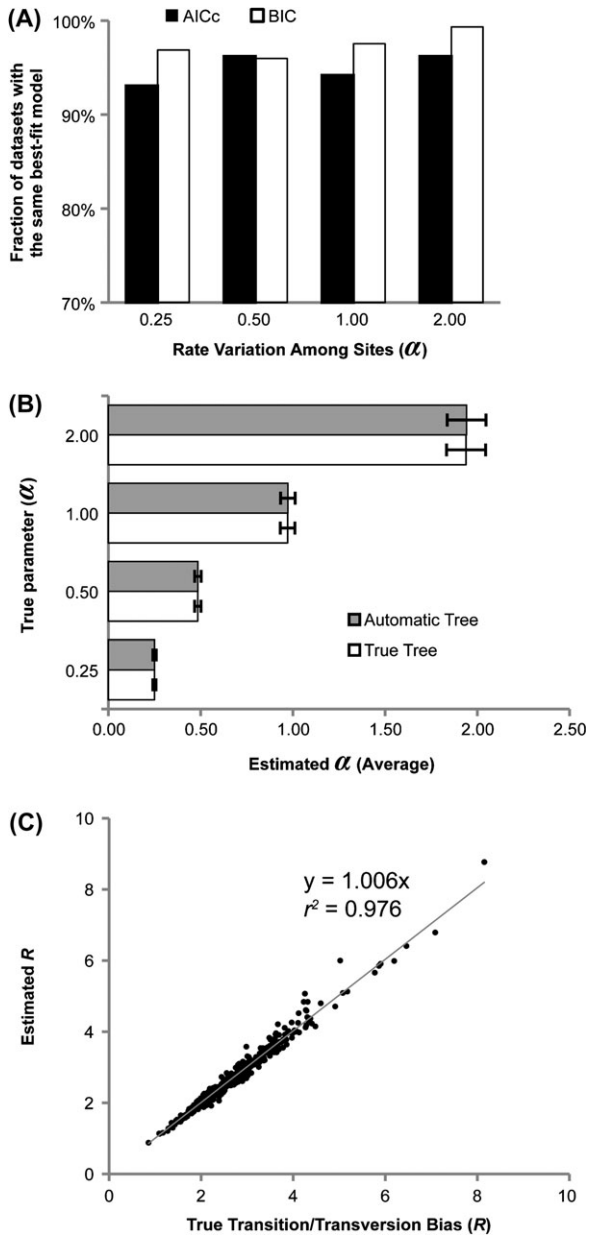

FIG. 2. Comparison of the best-fit model identified by using automatically generated and true trees for 1,792 computer simulated 66-sequence data sets. (*A*) The percentage of datasets for which the use of an automatically generated tree produces the same best-fit model as does the use of the true tree. Results are shown from datasets simulated with four different values of the gamma parameter (α) for rate variation among sites. (*B*) The estimates of α when using the automatically generated trees (filled bars) and the true tree (open bars). The average α and ± 1 standard deviation are depicted on each bar; 10 discrete Gamma categories were used. (*C*) The relationship of true and estimated transition–transversion ratio, *R*, when using automatically generated trees for data simulated with $\alpha = 0.25$. The value of *R* becomes 0.5 when the transition–transversion rate ratio, κ, is 1.0 in Kimura's two-parameter model. The slope of the linear regression was 1.005, with the intercept passing through the origin ($r^2 = 0.98$). Using the true tree, slope and r^2 values were 1.007 and 0.98, respectively. The absolute average difference between the two sets of estimates was 0.2% (maximum difference = 5.2%). Similar results were obtained for data simulated with $\alpha = 0.5$, 1.0, and 2.0.

found this approach to produce better initial estimates when there are many insertions–deletions and missing data in the sequences. After this procedure, the user-selected data subset is restored and used in all subsequent calculations.

By default, MEGA5 conducts an NNI search starting with the initial tree, such that the alternative trees differ in one branching pattern. One can expand the search space by using the CNI option in which alternative trees with two branches differences are evaluated (e.g., Nei and Kumar 2000, p. 126–127). In each case, ML values are computed for all the alternative trees produced by the branch swapping and all the branch swaps identified to increase the ML value are made simultaneously. If several single rearrangements are found to improve ML values for any branch, we choose the rearrangement that leads to the highest improvement in the ML value. We do not skip any branch swaps as long as it improves the ML value. In order to make major computational time savings, we do skip the evaluation of alternative topologies generated by rearrangements involving branches whose lengths are more than three times longer than their approximate standard errors. We use the second derivative of the ML score to generate approximate standard errors (Edwards 1972) during the branch length optimizations. Therefore, starting with systematic topological rearrangements of the initial tree, we discover trees with a higher ML value. These trees are subjected to new rounds of rearrangements, and this iterative process continues until no trees with greater likelihood can be found.

We tested the performance (time and accuracy) of the NNI and CNI searches in MEGA5 by means of computer-simulated data sets containing 66 sequences (see Materials and Methods). We compared the time taken to complete these heuristic searches with each other and with those needed by PhyML version 3.0 (Guindon et al. 2010) and RaxML version 7.0 (Stamatakis 2006). Results showed that, on average, a CNI search requires twice the time of an NNI search in MEGA5 (fig. 3*A*). Speeds of the MEGA5-NNI and MEGA5-CNI searches were similar to RaxML7-Mix and RaxML7-G, respectively. But, they were faster than PhyML3-NNI and PhyML3-SPR searches, respectively (fig. 3*A*). Similar trends were observed for another simulated dataset in which an increasingly larger number of sequences were analyzed (fig. 3*B*; 20–765 sequence data sets). For these data, the ML heuristic time increase shows a power trend with the increasing number of sequences (fig. 3*B*). It is important to note that the RaxML will be faster than MEGA5 if the user's machine is equipped with multiple processor and/or multicore CPUs because parallel versions of MEGA5 are yet to be implemented.

Even though different programs and search options show large differences in computational times, the average accuracies of the inferred ML trees were found to be rather similar. The accuracy difference is less than 3% for the data sets containing 66 sequences (fig. 4*A*) and 765 sequences (fig. 4*B*). Therefore, ML methods in MEGA5 appear to be comparable to other widely used ML implementations in terms of computational time and phylogenetic accuracy. In these simulations, we also compared the estimates of ML values generated by MEGA5 and

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

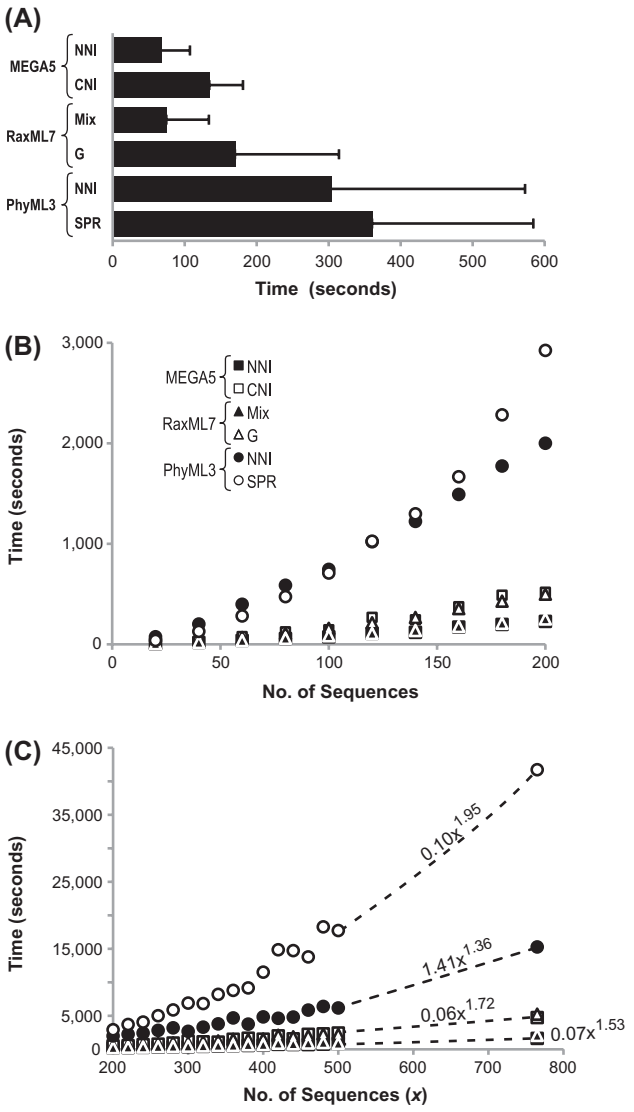

FIG. 3. Comparison of the computational speed of ML heuristic searches. (*A*) Average time taken to complete MEGA5 (NNI and CNI), RaxML7 (G and MIX), and PhyML3 (NNI and SPR) heuristic searches for 1,792 simulated data sets containing 66 sequences each. Bars are shown with ±1 standard deviation. Three data sets were excluded from PhyML3 calculations, as the NNI search failed. (*B*, *C*) Scatter plots showing the time taken to search for the ML tree for alignments that contain 20–200 and 200–765 sequences of 2,000 base pairs. The power trend fits are indicated for PhyML3 and MEGA5 ($r^2 > 0.98$ in all cases). For direct comparisons, all analyses were conducted by using 4 discrete categories for the Gamma distribution and a GTR model of nucleotide substitution (see Materials and Methods for simulation procedures, analysis descriptions, and computer hardware used). G, GTRGAMMA with four discrete Gamma categories; MIX, mixed method of using CAT and GAMMA models.

RaxML7 for the true tree and found them to differ less than 0.1% over all simulation cases.

Inference of Ancestral States and Sequences

MEGA5 now provides inferences of ancestral states and sequences using the empirical Bayesian method (fig. 5). Given a phylogenetic tree, branch lengths are estimated under a user-selected model of nucleotide or amino acid

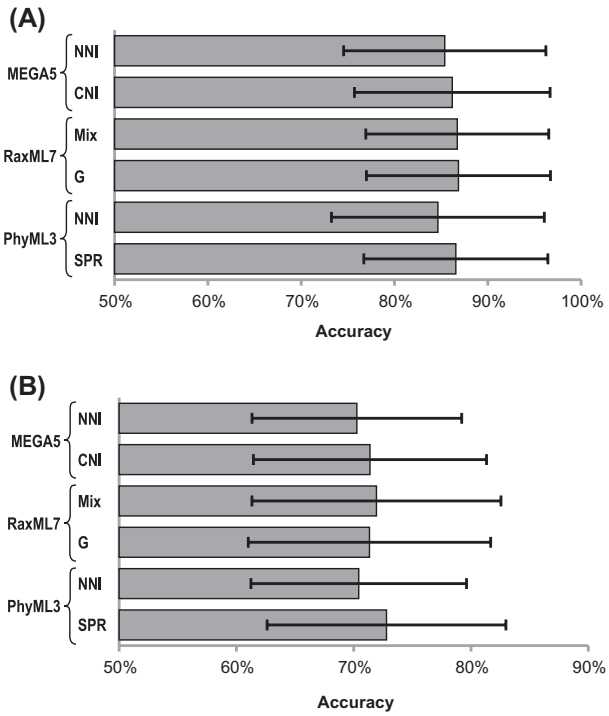

FIG. 4. Accuracies of heuristic ML trees produced by MEGA5, RaxML7, and PhyML3 programs. Shown are the proportions of interior branches (tree partitions) inferred correctly, along with ±1 standard deviation, for simulated data sets containing (*A*) 66 sequences and (*B*) 765 sequences. G, GTRGAMMA with four discrete Gamma categories; MIX, mixed method of using CAT and GAMMA models.

substitution and the Bayesian posterior probabilities are generated for each possible ancestral state assignment for each node (Yang et al. 1995). With this addition, users can now explore ancestral sequences inferred using maximum parsimony and ML methods in MEGA5. However, the latter is often preferable because it helps investigators to distinguish among multiple equally likely (most parsimonious) assignments by using the posterior probabilities for each possible nucleotide or amino acid assignment. Furthermore, it is expected to produce more accurate results at positions that have undergone multiple substitutions over the whole tree or in specific lineages (Nei and Kumar 2000). These ancestral states, along with posterior probabilities, can be exported in multiple formats for individual positions and for all complete ancestral sequences. However, note that the reconstructed ancestral sequences are not real observed data and may involve systematic biases and random errors, especially for the highly variable positions, so caution should be exercised if they are to be used in further statistical analysis.

Position-by-Position Evolutionary Rates

For both nucleotide and amino acid sequence data, users can estimate relative rates of molecular evolution position-by-position in MEGA5. Users select the number of discrete categories to approximate the Gamma distribution, specify whether or not to model invariant positions, and choose a nucleotide or amino acid substitution model. As mentioned earlier, they can use an automatically generated

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

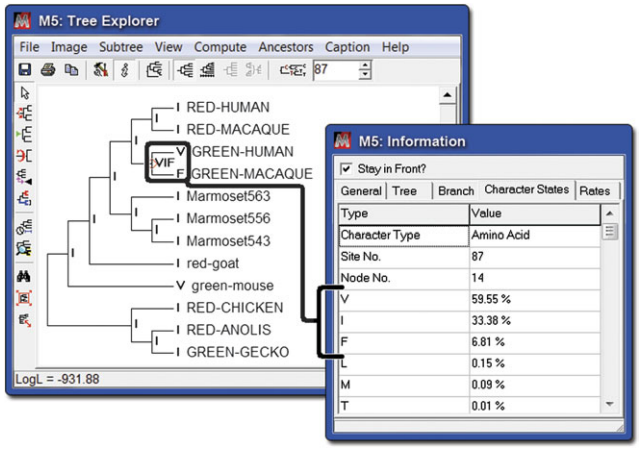

Fig. 5. Position-specific inferred ancestral states in a primate opsin phylogeny and the posterior probabilities of alternative amino acids at that position. See MEGA5 Examples directory for the data file and Nei and Kumar (2000, p. 212–213) for a description of the data. This figure is available in color online and in black and white in print.

topology, but it should be done carefully because the site-specific estimates of the evolutionary rate may depend on the evolutionary tree used (see Mayrose et al. 2005). No information on sequence divergence times is needed for estimating relative rates of evolution over sites, where all individual relative rates are scaled such that the average relative rate over all positions is equal to 1. This means that positions showing a relative rate less than 1 are inferred to be more highly conserved than the average conservation of sites in the alignment. Whenever available, these results are automatically exported directly to statistical analysis software, including Microsoft Excel, which can be used to generate sequence-wide profiles and conduct further analyses.

ML Molecular Clocks and Linearized Tree

In addition to Tajima's nonparametric test of molecular clock for three sequences (Tajima 1993), we have now added a likelihood ratio test of the molecular clock where the ML value for a given tree assuming the rate uniformity among lineages is compared with that without the assumption. In the output, primary information along with the *P* value of rejecting the null hypothesis of equal rates under a χ^2 distribution is presented. This test is expected to reject the null hypothesis when applied to data sets containing many sequences or long sequences as the strict equality of evolutionary rates among lineages is frequently violated. On the other hand, the estimates of branch lengths, and thus interior node depths, in a tree obtained under the assumption of a molecular clock can be useful to generate a rough idea about the relative timing of sequence divergence events (e.g., Takezaki et al. 1995). Of course, such estimates should be used cautiously.

Therefore, MEGA5 now provides estimates of ML branch lengths by assuming equal evolutionary rate among lineages. With this addition, users can now produce linearized trees using pairwise distances as well as the ML method. One can manually calibrate the molecular clock by setting the divergence time for any one node in the tree, which produces divergence times for all other nodes in the tree. For these divergence times, MEGA5 calculates approximate confidence intervals from the variance of the node height computed using the curvature method (e.g., Schrago 2006). Note that this procedure may underestimate the variance considerably due to the violation of the assumed clock constancy. The estimated node heights may be biased because of this reason, as well. So, the confidence intervals presented are not appropriate for hypothesis testing.

Usability Enhancements

We have also introduced many improvements to enhance MEGA's usability. First, MEGA5's central user interface has now become activity driven where a launch bar provides direct access to the growing suite of tools according to the type of analysis needed through the "Action Bar" (fig. 6). Once a user selects what they wish to compute, MEGA5 prompts for a data file to use and the methods and data subsets to employ. This wizard-style layout will make MEGA5 easier for beginners and expert users alike. In this spirit, we have now added native support for the widely used FASTA file format for sequence data, and sequence data can now be aligned using the MUSCLE software, which is very fast and accurate for data sets containing a large number of sequences (Edgar 2004). Because MEGA now accepts user trees for heuristic searches, for molecular clock tests, and for ancestral sequence reconstruction, we have included a tree topology editor that is useful for creating trees and editing existing topologies by using drag-and-drop of branches.

Operating Systems and Platforms

In a recent survey of long-term MEGA users, we have found that both Mac OS X and Linux platforms are used by a substantial number of researchers (one out of four). Therefore, we have been optimizing the use of MEGA5 on the Mac OS X and Linux platforms. For Mac OS X, we have now developed a custom installer that bundles MEGA5 and the WINE software so that the installation of MEGA5 is as simple as installing native Mac applications. WINE is a translation layer capable of running native Windows applications on POSIX compatible operating systems, such as Mac OS and Linux, and has two major advantages over using an emulation layer (i.e., virtualization software). First, by not using virtualization, users are not required to purchase a license for an additional operating system. Second, installation is simplified as there is no need to create and/or install an operating system disk image. As a result, Mac OS X users are able to use MEGA5 as seamlessly as if they were operating it on the Windows platform for which MEGA5 was originally developed. Similar provisions have been made for the use of MEGA5 on the Linux platform. In our tests, we found that calculations in MEGA5 on Mac OS X and Linux are $<5\%$ slower than Windows. This difference is rather small because all calculations via WINE are executed directly on the CPU like any other

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

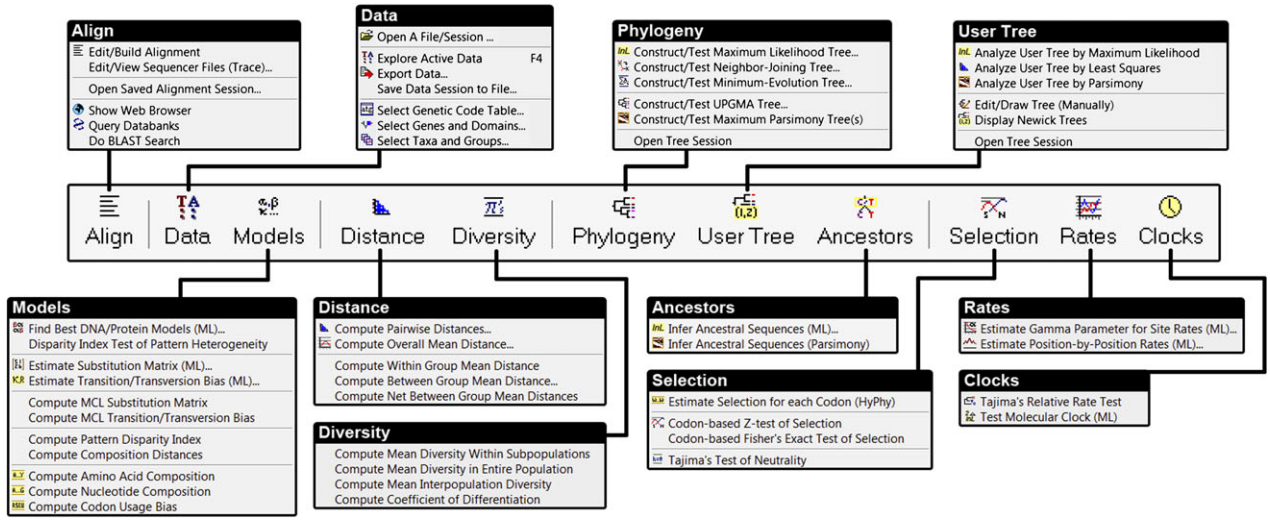

FIG. 6. The MEGA5 "Action Bar" and associated action menus. This figure is available in color online and in black and white in print.

native application in Mac OS X and Linux. In contrast, the MEGA5 user interface is rendered via emulation by WINE, which can sometimes result in a slowdown when drawing on the screen. But, this is becoming less noticeable with contemporary CPUs that are extremely fast. This enhancement is likely to make MEGA5 more useable for a greater number of researchers.

Conclusion

In summary, MEGA5 now provides analysis tools for three major types (ML, MP, and evolutionary distances) of statistical methods of molecular evolution (table 1 and fig. 6). These facilities not only make MEGA useful for more researchers but also enable researchers to evaluate the robustness of their results by comparing inferences from multiple methods under a variety of statistical models. In the future, we will continue to develop MEGA with a focus on implementing faster algorithms for phylogenetic inference, integrating more third party tools, and upgrading the computing core to use multicore and distributed computing effectively. As always, all versions of MEGA are available free of charge from http://www.megasoftware.net.

Materials and Methods

We generated two sets of nucleotide sequence data by using computer simulations. In one, a 66-taxa tree representing the phylogenetic relationships among mammals was used (see fig. 1 in Rosenberg and Kumar 2001). We simulated DNA evolution for 448 hypothetical genes along this tree, each with an independent set of evolutionary parameter values (base frequencies, sequence length, mean evolutionary rate, and transition–transversion rate ratio) estimated from the real sequence data (Rosenberg and Kumar 2003). For each set of evolutionary parameters (448 different sets), the branch lengths of the model tree were estimated using the corresponding evolutionary rate. Sequence alignments were generated under the Hasegawa–Kishino–Yano (HKY) (Hasegawa et al. 1985) model of nucleotide substitution at four different levels of rate variation among sites ($\alpha = 0.25, 0.5, 1.0$, and 2.0) that were implemented during computer simulations via a discretized gamma distribution with a very large number of categories. This resulted in a total of 1,792 alignment sets.

We also generated DNA sequence alignments containing 20–765 taxa, which were based on the corresponding sized trees derived from a master phylogeny of 765 taxa (see supplementary fig. S1, Supplementary Material online, in Battistuzzi *et al.* [2011]). This master phylogeny was obtained by pruning taxa and groups from the tree of 1,671 families in the Timetree of Life (Hedges and Kumar 2009), such that the final tree was strictly bifurcating. The resultant tree of 765 taxa was scaled to time and spanned 4.2 billion years of evolution. This master topology was subsampled to produce model trees used to generate the sequence alignments containing varying number of taxa (20, 40, 60, . . ., 500), with one set containing all 765 taxa. Sequences were simulated using SeqGen (Rambaut and Grassly 1997) under the HKY (Hasegawa et al. 1985) model of nucleotide substitution with a G + C content of 48% and a transition–transversion rate ratio of 1.05, which were estimated from an alignment of small subunit rRNA sequences from 800 taxa of animals, fungi, plants, and archaebacteria. In order to make the evolutionary rate heterogeneous among tip and internal lineages, rates were varied randomly by drawing them from a uniform distribution with boundaries ±5% of the expected rate in each branch independently. We used substitution rates of 0.025, 0.050, and 0.100 per base pair per billion years to establish branch lengths. In total, 530 data sets were generated in this way and the results are presented in the main text. We also conducted 290–765 taxa simulations in which sequences evolved four times faster (0.4 substitutions per site per billion years) and found the differences between methods were very similar to those reported here (results not shown).

A benchmark comparison of ML phylogenetic inference between MEGA5, RAxML7, and PhyML3 was performed for all simulated datasets by collecting the computational and phylogenetic performance of these programs, including

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

execution time (in seconds), the estimate of a gamma shape parameter, ML values, and topological accuracy. Because Windows is MEGA5's native operating system, Windows executables were used for PhyML (version 3.0) and RAxML (version 7.04). All analyses were conducted on computers with identical hardware (Intel Q8400 2.66 GHz Quad Core processor and 6 GB RAM) and operating systems (64-bit Windows 7 Enterprise Edition). For direct comparison, each program was executed serially in a single thread of execution with one core utilized per dataset.

In order to generate comparable results on time and accuracy, we used identical substitution models and discrete gamma options across all programs. Because the fastest heuristic search in RaxML, MIX, assumes a general time reversible (GTR) model with four discrete gamma rate categories, we used these options in all cases, unless noted otherwise. For all three programs, analyses were conducted using the automatically generated initial trees and selecting the default options. And, heuristic searches starting with the initial trees were conducted with two different levels of branch rearrangements: quick searches (Nearest Neighbor Interchange [NNI] for MEGA5 and PhyML and MIX for RaxML) and slow searches (Close Neighbor Interchange [CNI] for MEGA5, Subtree–Pruning–Regrafting [SPR] for PhyML, and GTRGAMMA for RaxML). The accuracy of phylogenetic tree of n taxa was estimated from the topological distance (d_T) between the inferred tree and the true topology was given by $(n - 3 - \frac{1}{2}d_T)/(n - 3)$.

Acknowledgments

We thank the colleagues, students, and volunteers who spent countless hours testing the early release versions of MEGA5. Many facets of the user-interface design benefited from the extensive comments of the members of our laboratories and of users at large. We also thank Mr Paul Billing-Ross for his help with computer simulations and Ms Carol Williams for editorial support. The MEGA software project is supported by research grants from National Institutes of Health to S.K. and M.N.

References

Alfaro ME, Huelsenbeck JP. 2006. Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst Biol*. 55:89–96.

Battistuzzi FU, Billing-Ross P, Paliwal A, Kumar S. 2011. Fast and slow implementations of relaxed clock methods show similar patterns of accuracy in estimating divergence times. *Mol Biol Evol*. 28:2439–2442.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5:113.

Edwards AWF. 1972. Likelihood; an account of the statistical concept of likelihood and its application to scientific inference. Cambridge (UK): Cambridge University Press.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17:368–376.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Fitch WM. 1986. An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. *Prog Clin Biol Res*. 218:149–159.

Fitch WM, Margoliash E. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem Genet*. 1:65–71.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14:685–695.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59:307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

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

Hedges SB, Kumar S. 2009. Discovering the timetree of life. In: SB Hedges, S Kumar, editors. The timetree of life. New York: Oxford University Press. p. 3–18.

Hurvich CM, Tsai C-L. 1989. Regression and time series model selection in small samples. *Biometrika* 76:297–307.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 8:275–282.

Kumar S, Dudley J. 2007. Bioinformatics software for biologists in the genomics era. *Bioinformatics* 23:1713–1717.

Kumar S, Nei M, Dudley J, Tamura K. 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform*. 9:299–306.

Kumar S, Tamura K, Nei M. 1994. MEGA: molecular evolutionary genetics analysis software for microcomputers. *Comput Appl Biosci*. 10:189–191.

Mayrose I, Mitchell A, Pupko T. 2005. Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *J Mol Evol*. 60:345–353.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. Oxford: Oxford University Press.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol*. 53:793–808.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*. 13:235–238.

Ripplinger J, Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? *Syst Biol*. 57:76–85.

Ripplinger J, Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. *Mol Biol Evol*. 27:2790–2803.

Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc Natl Acad Sci U S A*. 98:10751–10756.

Rosenberg MS, Kumar S. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Mol Biol Evol*. 20:610–621.

Saitou N, Nei M. 1987. The Neighbor-Joining Method—a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406–425.

Schrago CG. 2006. An empirical examination of the standard errors of maximum likelihood phylogenetic parameters under the molecular clock via bootstrapping. *Genet Mol Res*. 5:233–241.

Schwarz G. 1978. Estimating the dimension of a model. *Ann Stat*. 6:461–464.

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024

Shoemaker JS, Fitch WM. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol Biol Evol*. 6:270–289.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21:456–463.

Swofford DL. 1998. PAUP*: phylogenetic analysis using parsimony (and other methods). Sunderland, MA: Sinauer Associates.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135:599–607.

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol*. 12:823–833.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 10:512–526.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A*. 101:11030–11035.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39:306–314.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.

Downloaded from https://academic.oup.com/mbe/article/28/10/2731/973375 by guest on 24 April 2024