An Amino Acid Substitution-Selection Model Adjusts Residue Fitness to Improve Phylogenetic Estimation

Huai-Chun Wang,*[1,2] Edward Susko,[1,2] and Andrew J. Roger[2,3]

[1]Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

[2]Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia, Canada

[3]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

***Corresponding author:** E-mail: huaichun.wang@dal.ca.

Associate editor: Tal Pupko

Abstract

Standard protein phylogenetic models use fixed rate matrices of amino acid interchange derived from analyses of large databases. Differences between the stationary amino acid frequencies of these rate matrices from those of a data set of interest are typically adjusted for by matrix multiplication that converts the empirical rate matrix to an exchangeability matrix which is then postmultiplied by the amino acid frequencies in the alignment. The result is a time-reversible rate matrix with stationary amino acid frequencies equal to the data set frequencies. On the basis of population genetics principles, we develop an amino acid substitution-selection model that parameterizes the fitness of an amino acid as the logarithm of the ratio of the frequency of the amino acid to the frequency of the same amino acid under no selection. The model gives rise to a different sequence of matrix multiplications to convert an empirical rate matrix to one that has stationary amino acid frequencies equal to the data set frequencies. We incorporated the substitution-selection model with an improved amino acid class frequency mixture (cF) model to partially take into account site-specific amino acid frequencies in the phylogenetic models. We show that 1) the selection models fit data significantly better than corresponding models without selection for most of the 21 test data sets; 2) both cF and cF selection models favored the phylogenetic trees that were inferred under current sophisticated models and methods for three difficult phylogenetic problems (the positions of microsporidia and breviates in eukaryote phylogeny and the position of the root of the angiosperm tree); and 3) for data simulated under site-specific residue frequencies, the cF selection models estimated trees closer to the generating trees than a standard Γ model or cF without selection. We also explored several ways of estimating amino acid frequencies under neutral evolution that are required for these selection models. By better modeling the amino acid substitution process, the cF selection models will be valuable for phylogenetic inference and evolutionary studies.

Key words: **selection, amino acid substitution, maximum likelihood, site-specific frequencies, mixture model, molecular phylogenetics.**

Article

Fast Track

Introduction

Markov models of protein sequence evolution usually use an empirical rate matrix, Q, of instantaneous rates of exchange between amino acids to compute the probability of transition from one amino acid to another over any given evolutionary distance. The rate matrix Q is determined from a large database, and a number of such matrices exist including the PAM, JTT, WAG, and LG matrices (Dayhoff et al. 1978; Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008). These empirical rate matrices are derived in such a way as to be time reversible: $\pi_i Q_{ij} = \pi_j Q_{ji}$, where π_j is the stationary frequency of the amino acid j. A rate matrix and its corresponding π_j profile are often combined with a discrete Γ rate mixture model to account for the rate heterogeneity among the alignment columns.

The stationary frequency π_j is approximately the same as the frequency of the amino acid j over the large number of protein data sets that were used to derive the rate matrix. In many cases, these are not very close to the observed frequencies for the data set of interest. To adjust for this, a variant of this scheme, termed +F in phylogenetic models, sets the stationary amino acid frequencies to those observed in the data under analysis (Cao et al. 1994). Given the rate matrix, Q, for the empirical model, the usual adjustment for data set frequencies first determines what is commonly referred to an exchangeability matrix, R, through $R_{ij} = Q_{ij}/\pi_j$. The rate matrix for the +F model is then taken as $Q_{ij} \propto R_{ij}\, \pi_j^{(D)}$, where $\pi_j^{(D)}$ is the frequency of amino acid j for the data set and the row sums in Q are zeros. Here, as throughout the article, the constant of proportionality is determined from the constraint that $-\sum \pi_j^{(D)} Q_{jj} = 1$, which ensures that edge lengths are interpretable as expected numbers of substitutions per site. The +F method generally fits the data better than the model that uses the equilibrium frequencies of the protein substitution model and is widely used.

The foregoing +F method gives stationary frequencies equal to the data set frequencies while maintaining time reversibility. However, it is not the only way to manipulate the rate matrix to satisfy these conditions. For example, the

Mol. Biol. Evol. 31(4):779–792 doi:10.1093/molbev/msu044 Advance Access publication January 16, 2014

generalized weighted frequency (+gwF) model of Goldman and Whelan (2002) gives stationary frequencies equal to the data set frequencies but multiplies the exchangeabilities by the ratio of the frequencies of the target amino acids (π_j) to the source amino acids (π_i):

$$Q_{ij} \propto \frac{\pi_j^{1-f}}{\pi_i^{f}} R_{ij}$$

Here, f is a parameter that controls the relative contributions of the frequencies of the source and target residues and was estimated through maximum likelihood (ML). When $f = 0$, Q_{ij} is reduced to the standard +F variant of the Q matrix.

The +gwF model was found to fit 70% of the 182 test data sets better than the +F model (Goldman and Whelan 2002). Both +F and +gwF methods provide somewhat arbitrary ways of adjusting for different amino acid frequencies in a given data set. Moreover, these models do not explicitly adjust for selection in terms of the relative fitness of the various residues at sites. Different sites in a protein are selected to have different structural or functional roles resulting in different substitution patterns at individual positions. For instance, isoleucine is commonly found in buried β-stranded environments in which it is often substituted by valine and leucine. However, in a context where the position is functionally conserved with isoleucine, it is less likely that isoleucine can be substituted by these residues (Sjölander et al. 1996). Similarly, phenylalanine seen in a context that requires an aromatic residue is often found to be substituted by tyrosine or tryptophan, whereas under a context requiring a large nonpolar residue, the substitutions with aliphatic or other large residues are common. This site-specific amino acid substitution pattern is not modeled by the equilibrium amino acid frequencies from the protein model nor by the +F or +gwF variants.

Halpern and Bruno (1998) first applied population genetics theory to derive a rate matrix for the codon substitution process, which combines the nucleotide mutation probabilities with the fixation probabilities of the codons at individual sites. Their model was further developed to detect selective strengths on synonymous codon usage (Yang and Nielson 2008), to estimate the distribution of selection coefficients from phylogenetic data (Tamuri et al. 2012), and to study the heterogeneity of amino acid fitness profiles among sequence sites (Rodrigue et al. 2010; Rodrigue 2013). Holder et al. (2008) explicitly studied the impact of the Halpern–Bruno (HB) model on phylogenetic estimations and found standard protein models had difficulty recovering the correct phylogenies when the data were simulated for the site-specific amino acid frequencies under the HB model especially when the simulated data were divergent and sequences were short. Other early work also introduced fitness parameters in amino acid sequences to study protein evolution, but the fitness was not modeled in terms of the mutation rates and amino acid frequency profiles and was not site specific (Dimmic et al. 2000). In this study, we introduce new substitution-selection models for protein sequences combined with mixture models of class frequency profiles of amino acid sites (Wang et al. 2008) (extended to 9 and 20 site classes). Our tests on empirical and simulated data show that these amino acid selection models can improve the ML estimation of phylogenies.

New Approaches

Following Halpern and Bruno (1998) as well as Yang and Nielson (2008) and Rodrigue et al. (2010), we define the fitness of amino acid j in a protein sequence alignment as the log of the ratio of the frequency of the amino acid (π_j) to the frequency of the same amino acid under no selection (π_{j_0}): $F_j = \log \frac{\pi_j}{\pi_{j_0}}$. It can be shown, based on population genetics theory (Kimura 1962), that the population scaled selection coefficient $S_{ij} = 2N \quad s_{ij} = \log \frac{\pi_j}{\pi_{j_0}} - \log \frac{\pi_i}{\pi_{i_0}} = F_j - F_i$ (see also Rodrigue et al. 2010). The rate matrix for residue changes under selection, $Q_{ij}^{(s)}$, is a product of a rate matrix under no selection, $Q_{ij}^{(0)}$, and the probability of fixation of mutant j given i, f_{ij}.

$$Q_{ij}^{(s)} \propto Q_{ij}^{(0)} f_{ij} = Q_{ij}^{(0)} \frac{F_j - F_i}{1 - e^{-(F_j - F_i)}}$$
$$= Q_{ij}^{(M)} \frac{1 - e^{-(F_j^{(M)} - F_i^{(M)})}}{F_j^{(M)} - F_i^{(M)}} \frac{F_j - F_i}{1 - e^{-(F_j - F_i)}}$$

where $Q^{(M)}$ is any standard amino acid substitution rate matrix such as LG and $F_j^{(M)}$ satisfies the F_j formula but with π_j replaced by the equilibrium frequency $\pi_j^{(M)}$ from the rate matrix $Q^{(M)}$. Three simple π_{j_0} profiles are proposed and tested in this article: the first one results from equal codon frequencies of all 61 sense codons; the second is also based on equal codon frequency but adjusts for nucleotide compositional bias manifested by the ratio of amino acids encoded by GC-rich codons (G, A, R, and P) to those encoded by AT-rich codons (F, Y, M, I, N, and K); and the third one is determined from the nucleotide frequencies at the third codon positions. We call the selection models under the three π_{j_0} profiles as selection 1, selection 2, and selection 3, respectively (sel1, sel2, and sel3 for short). See the Materials and Methods section for additional details about the derivation of $Q_{ij}^{(s)}$ and the estimation of π_{j_0}.

Given $Q_{ij}^{(s)}$, the Q matrix adjusted for selection effects, one can estimate the phylogenies under the standard ML framework (Felsenstein 1981). Furthermore, we combine a class frequency mixture model (Wang et al. 2008) with the selection model to take into account site-specific substitution patterns. In addition to the original four site classes (cF4) introduced in Wang et al. (2008), we utilized two sets of previously published amino acid frequency profiles: the nine component profiles of Sjölander et al. (1996) and the 20 component profiles of Quang et al. (2008). The corresponding cF models using these profiles are referred to as cF9 and cF20, respectively, to distinguish them from the original cF4 model. Each of the three mixture models contains an F component (the average amino acid frequency of the alignment). The F component in the cF models may be adjusted for selection effects as introduced earlier; the resulting cF models with selection are called cF4sel, cF9sel, and cF20sel, respectively, where only the F component is under selection. The base model with a single rate matrix and +F component is called Fsel, when selection is implemented. Moreover, as defined earlier the various selection models can be sel1, sel2, or sel3 (e.g., the cF4sel model can be cF4sel1, cF4sel2, and cF4sel3) depending on how the neutral amino acid frequency profile (π_{j_0}) is estimated.

Results and Discussion

Model Test: 21 Empirical Data Sets

We first applied the Fsel1 and Fsel2 selection models to the 21 protein family data sets and calculated the maximum log-likelihood score (ln *L*) of the data sets for the trees pre-estimated under a standard model without adjusting for selection (the base model LG + F + Γ). For these models, sel1 determines the neutral amino acid frequencies implied by equal codon frequencies, whereas sel2 further adjusts for the GARP/FYMINK ratio. Both Fsel1 and Fsel2 obtained larger ln *L* scores in 18 data sets than the base model. Because the Fsel1 model does not increase the number of free parameters relative to the base model, the likelihood gains indicate that Fsel1 fits better than the base model without selection in the majority of the data sets. Fsel2 has one more parameter than the base model. Still in all but three data sets that both Fsel1 and Fsel2 had lower corrected Akaike information criterion (AIC$_c$) than the base model. Comparing Fsel2 and Fsel1, the ln *L* scores were higher in 17 out of the 21 data sets for Fsel2, and two of them gave lower AIC$_c$ scores for Fsel2, or alternatively, were significant using a likelihood ratio (LR) test ($\chi^2_{1,0.05} = 3.841$).

Next we applied the cF mixture models (cF4, cF9, and cF20) and their corresponding selection models to estimate the ML scores for the 21 data sets. For comparison, an empirical profile mixture model (CAT-C20 or CAT20; Quang et al. 2008) was also used to analyze the data sets. Figure 1 right panel shows a boxplot of the log-likelihood difference (Δln *L*) between the various models and the base model, and figure 1 left panel shows a boxplot of per site AIC$_c$ differences (base − test model) for the 21 data sets. Compared with the base model, the likelihood increase for all data sets was very significant for the three cF mixture models (cF4, cF9, and cF20) and their counterparts that included selection on the F component (LR test $P < 0.001$). They also had smaller AIC$_c$ scores than the base model for all data sets. Because the three cF models are not nested, the AIC$_c$ were used to compare cF4, cF9, and cF20. Both cF9 and cF20 achieved smaller AIC$_c$ scores than cF4 for all data sets; cF20 had smaller AIC$_c$ scores than cF9 for 15 data sets, whereas cF9 had smaller AIC$_c$ scores in the remaining six data sets. All together this indicates the mixture profiles with more components of amino acid frequency vectors tend to improve model fit, consistent with the findings in Quang et al. (2008). CAT20, however, gave higher likelihoods than the base model in just about half of the data sets (12 out of 21) and the AIC$_c$ indicated it performed better than the base model in only 10 data sets. This agrees with Le et al. (2008), which found CAT20 was no better than the basic LG + F + Γ especially for unsaturated data sets. Adding a + F to the CAT20 model (CAT20 + F) did not improve performance. In fact the estimated weights for the F component were less than 0.05 for all but one data set. However, when the proportional exchangeabilities of the standard CAT20 model (Quang et al. 2008) were replaced by the LG exchangeabilities, a model we called CAT20 + LG model, estimated likelihoods were significantly better than under the base model for all data sets (data not shown). Therefore, it appears valuable to combine the C20 profiles with an empirical protein matrix (such as LG) instead of a protein matrix under a Poisson model.

The cFsel1 models have the same number of parameters as the corresponding cF models, so comparison of the likelihoods can be made directly. For instance, cF4sel1 gave larger likelihoods than cF4 in 17 data sets; both cF9sel1 and cF20sel1 models gave larger likelihoods than cF9 and cF20 in 15 data sets. Therefore, each of the cFsel1 models fit better than the corresponding cF model in the majority of the data sets. Comparing Msel2 with Msel1 for the same model M (M being cF4, cF9, or cF20), Msel2 gave higher ln *L*'s in most of the cases. For example, cF4sel2 showed higher ln *L* than cF4sel1 in 17 data sets; cF9sel2 had better ln *L* than cF9sel1 in 16 data sets; and cF20sel2 was better than cF20sel1 in 15 data sets. However, in all but one of the cases, the likelihood improvements were not sufficiently large to overcome the cost of an additional parameter in the sel2 models according to the AIC$_c$.

It should be noted in the above applications, the cFsel models, like the Fsel models, only had the F component (i.e., the average amino acid frequency of the data set) placed under selection. In principal, one can also put the cF components under selection using similar equations as equations (5) and (6) to adjust the *Q* matrices. We found this did not always increase model fit. For example, when we modified the cF20sel model, so that all 20 components and F were under selection, higher likelihood scores than in the cF20 model were obtained in only 10 data sets. By comparison, the cF20sel model with only F under selection achieved larger likelihood scores than cF20 in 15 data sets. The main reason is probably numerical instability, as the *Q* matrix adjustments involve in log ratio of the frequency vectors (eq. 4), and some of the cF profiles contain very small frequencies for certain amino acids, which make them sensitive to the sparseness of site pattern in the aligned sequence data. Therefore, the cF selection models introduced throughout the article, as implemented in QmmRAxML (version 2.0), are based on only the F component being placed under selection.

Figure 2A shows the log-likelihood gains from the eight test models (Fsel1, cF4, cF4sel1, cF9, cF9sel1, cF20, cF20sel1, and CAT20) relative to the base model plotted against a measure of the information content of the 21 data sets. The information content of a data set is taken as the product of the number of taxa, the aligned sequence length, and the length of the tree estimated under the base model. This quantity captures the total amount of sequence "change" in the data set. Positive relationships between the ln *L* gains and the information content of the data sets are evident in each

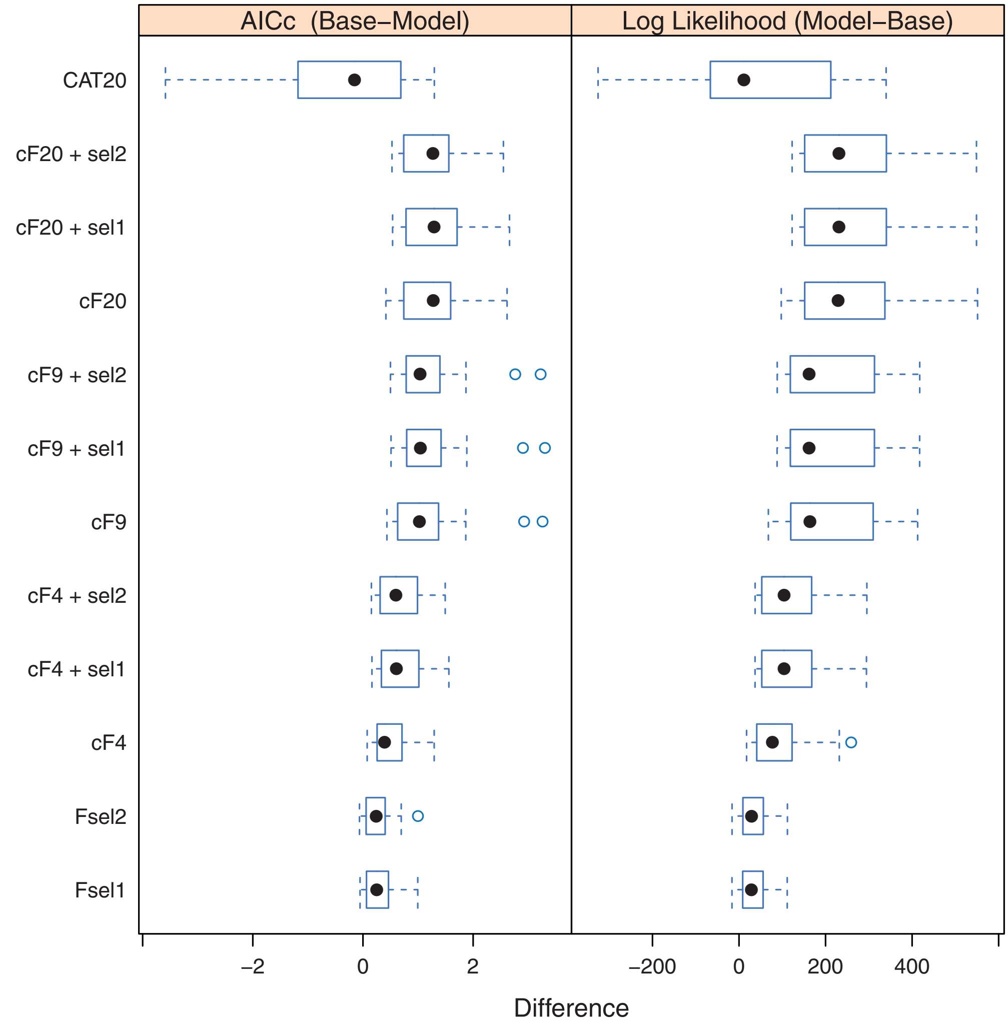

Fig. 1. Box plot (right panel) of the difference in ln L between the test models and the base model and box plot (left panel) of per site AIC$_c$ score differences (the base model − test model) for the 21 data sets. The base model is LG + F + Γ, and the test models are shown on the y axis.

model. Indeed, the correlation coefficients between the two variables are 0.46 (Fsel1), 0.42 (cF4), 0.53 (cF4sel1), 0.67 (cF9), 0.68 (cF9sel1), 0.62 (cF20), 0.63 (cF20sel1), and 0.24 (CAT20). The Fisher transformation test of correlation coefficient (r) differing from 0 gave P values less than 0.05 in all cases but cF4 and CAT20. This indicates that for most of the models, the greater the information content in a data set, the larger the expected increase in likelihood compared with the base model. Significant correlations between the ln L gains and the information content were also obtained for all sel2 models (Fsel2, cF4sel2, cF9sel2, and cF20sel2) relative to the base model. The likelihood increases of all cF selection models over corresponding cF models (e.g., cF4sel1 or cF4sel2 vs. cF4) were not significantly correlated with the information content ($r < 0.36$, $P > 0.05$).

The correlation between the average branch length in each data set and the ln L gains from the test models relative to the base model were all negative ($r = -0.49$ to -0.12) except in Fsel1 and Fsel2 ($r = 0.33$ in both cases). The Fisher transformation test for correlation gave P values greater than 0.05 in all cases except in CAT20 where a significantly negative correlation was observed ($r = -0.49$; $P < 0.05$). However, the ln L differences between the cF selection models and corresponding cF models (e.g., cF4sel1 vs. cF4) were all positively correlated with the average branch lengths with cF4sel1, cF4sel2, cF20sel1, and cF20sel2 being significant ($r = 0.44$ to 0.46, $P < 0.05$). Figure 2B plots the ln L differences in cF4sel1 and cF20sel1 (relative to cF4 and cF20, respectively) against the average branch length in each data set. Very similar results appeared if plotting the ln L gains in cF4sel2 and cF20sel2 (relative to cF4 and cF20, respectively) against the average branch lengths, suggesting that greater gains may be expected with more divergent sequence data.

Tree Estimation

Previously we showed that the cF model accounts for site-to-site heterogeneity in the substitution process and improves tree estimation by reducing the influence of artifactual long branch attraction (LBA) bias (Wang et al. 2008). For example, for the multigene microsporidia data (Brinkmann et al. 2005), the cF4 model estimated a higher likelihood score for the microsporidia-fungi (MF) clade tree than the microsporidia-archaea (MA) clan tree. The latter tree was often inferred under the conventional model employing a single rate

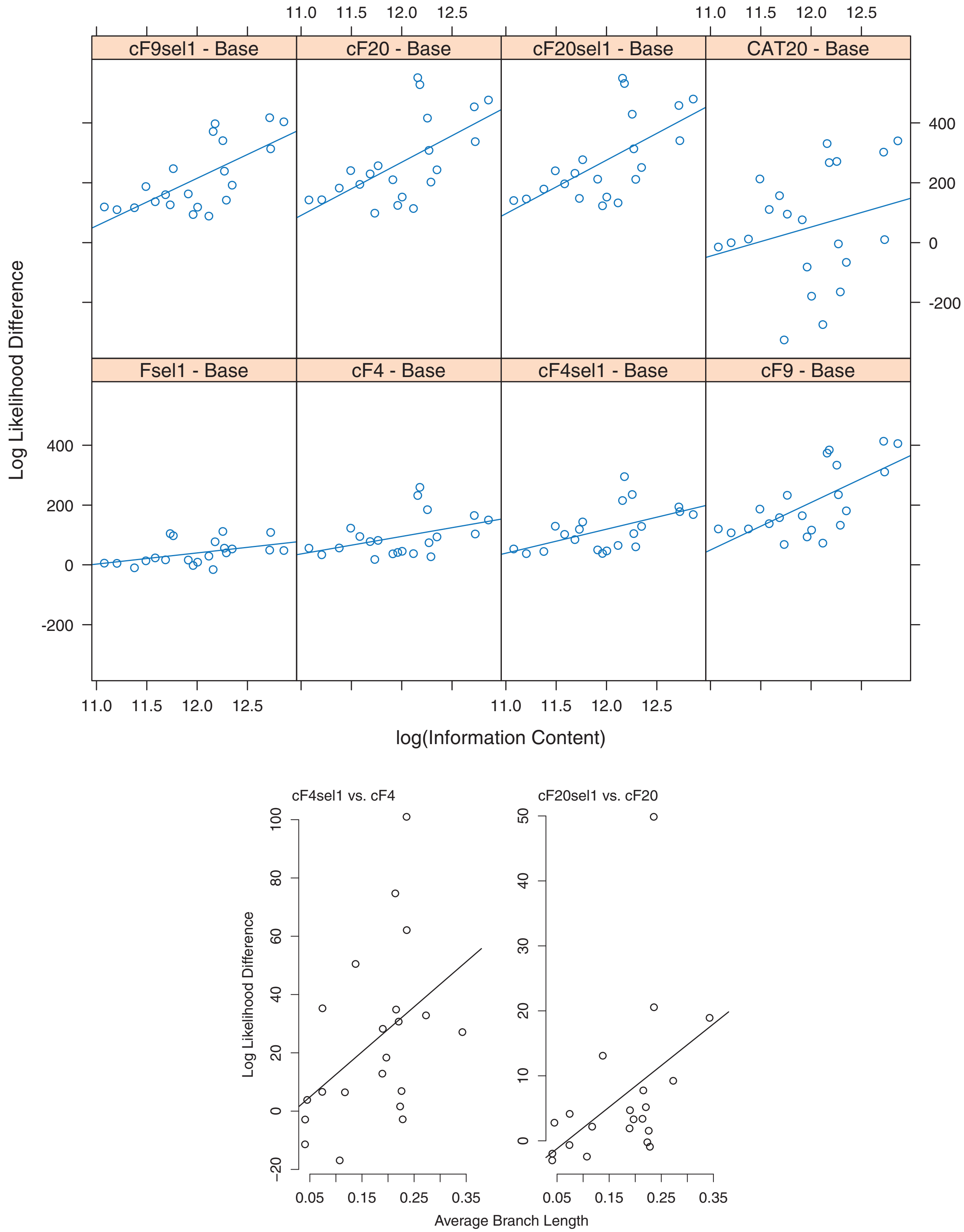

Fig. 2. (*A*) The ln *L* differences between the test models and the base model (LG + F + Γ) plotted against the log information content (taxa × sites × tree length). (*B*) The ln *L* differences between cF4sel1 and cF4 and between cF20sel1 and cF20 plotted against the average branch lengths. Each dot is for a protein data set.

matrix due to the LBA artifact, whereas an MF clade tree was inferred under more sophisticated models and methods that could reduce the LBA effect (Brinkmann et al. 2005). Here, we reanalyzed this data set under the cF4, cF9, cF20, and their selection models as well as the base model (LG + F + Γ), Fsel, and CAT20. The amino acid profile under no selection was based on equal codon frequencies with the GARP to FYMINK ratio adjusted. Table 1 lists the ln *L* scores for the MF tree and the ln *L* differences between the MF and MA trees for the no-selection and selection models. All cF models and cFsel2

Table 1. ML Estimations of the Microsporidia Sequence Data for Two Competing Trees.

Model[a]	ln L of the MF Tree	Δln L (MF − MA Trees)	P*(MA Tree)	AIC$_c$ per Site[b]	No. of Model Parameters	F Component Weight[c]
LG + F (base model)	−731688.61	−117.83	0.97	60.24	79	1.0
Base + sel2	−731002.66	−125.59	0.97	60.19	80	1.0
Base + cF4	−727360.77	7.72	0.44	59.89	83	0.67
Base + cF4 + sel2	−726869.82	3.84	0.23	59.85	84	0.68
Base + cF9	−721709.64	125.12	0.003	59.42	88	0.26
Base + cF9 + sel2	−721601.93	122.16	0.0	59.41	89	0.28
Base + cF20	−718325.89	156.13	0.0	59.14	99	0.27
Base + cF20 + sel2	−718215.49	154.56	0.0	59.14	100	0.28
Poisson + CAT20	−727747.48	135.86	0.0	59.92	98	NA

NOTE.—MF, microsporidia grouped with fungi; MA, microsporidia at the base of the eukaryotes close to the archaea species.

[a] sel2 is the selection model with the neutral amino acid profile estimated from equal codon frequency with GARP/FYMINK adjustment.

[b] Corrected AIC score was calculated based on the MA tree. Almost identical AIC per site scores were obtained if using the MF tree in the calculation.

[c] Weight for the F component was calculated based on the MA tree. Similar weight was obtained for each model if using the MF tree in the calculation. NA, the F component is not applicable for the model.

*P value from the KH test (Kishino and Hasegawa 1989). The null hypothesis is the indicated tree. Small P values suggest strong support for the other tree.

models estimated higher likelihoods for the MF tree, whereas the standard Γ model with or without selection estimated a higher likelihood for the MA tree. The CAT20 model also estimated larger likelihood for the MF tree. Table 1 also gives the P value of the KH test for the null tree (the MA tree) under each model based on the difference in the estimated site-wise log likelihoods between the null and alternate trees. The cF9, cF20, and corresponding selection models, as well as CAT20, strongly rejected the MA tree in favor of the MF tree, whereas the base model with or without selection supported the MA tree. Based on the corrected per-site AIC$_c$ scores in table 1, the cF20 and cF20sel models would be considered the best-fitting models.

Table 1 further reports the estimated weights for the F component in the cF4, cF9, cF20, and their corresponding cFsel2 models, based on the MA tree (similar results were obtained when the MF tree was used). The F component weights for cF9 and cF20 are comparable and less than the corresponding weight for cF4. This implies additional weight is being placed on the cF9 and cF20 mixture components, suggesting a substantial subset of sites are better modeled by some of the additional cF9 and cF20 frequency profiles. The F component weights for a given model are similar whether the model includes selection or not. Thus, including selection does not increase the number of sites attributed to the F component class. Better fits are likely due to selection better modeling the site patterns of the F component class.

A second phylogenomic data set we considered was the *Amborella* chloroplast genome data (Leebens-Mack et al. 2005). These authors found that *Amborella* alone at the base of angiosperm phylogeny (tree A) was supported by the amino acid sequence data, whereas *Amborella* + water lilies at the base of the flowering plants (tree B) was supported by the nucleotide data. Table 2 lists the ln L scores for tree A and the ln L difference between tree B and tree A under the various models with a chloroplast-specific amino acid substitution matrix (cpREV) (Adachi et al. 2000). The cpREV matrix was used here in place of the LG matrix used previously as it achieved higher ln L scores than LG or WAG for the same model. For instance, the ln L of tree A under cpREV + F + Γ was −175,584, much higher than −178,591 under LG + F + Γ. The KH test was unable to reject any of the two competing trees for each model listed in table 2. Tree B was only marginally better supported by covarion model-based ML estimations (Wang et al. 2007) and a more recent study based on a much larger data set (Goremykin et al. 2013). Table 2 shows the cF20 and cF20sel2 models were the best-fitting models according to AIC$_c$. Surprisingly, CAT20 had the highest AIC score and was higher than the base model. The reason may be that the CAT20 model uses a proportional rate matrix for exchangeabilities that does not fit the chloroplast data well.

Table 2 also lists the estimated weights for the F component in the cF and cFsel2 models. As in the case of the microsporidia data, the F component weights for cF9 and cF20 are less than the F weight for cF4. Comparing the F component weights under the cF models and those under the corresponding cF with selection models, they are very similar for cF4 ± sel2 and cF20 ± sel2, respectively. However, a large difference (0.15) in the estimated F component weights exists between cF9 and cF9sel2, which corresponds to a better, lower AIC$_c$ score in the cF9sel2 model (the per site AIC$_c$ difference is 0.05 between the two models).

Table 3 reports results for the protistan breviate data. Two competing trees were recently estimated (Brown et al. 2013): a BA tree, which put the breviate lineage sister to the apusomonads relative to the opisthokonts, was estimated under a standard LG + F + Γ base model, whereas an OA tree, which let the breviates splitting off before a sister clade of the opisthokonts and apusomonads, was estimated under a CAT-GTR + Γ model. All cF mixture models fit the data significantly better than the base model. The cF selection models had even higher ln L than corresponding models without selection. CAT20 fit the data better than the base, Fsel, cF4, and cF4sel2 models, but it was worse than the cF9, cF20, and their selection models. The per site AIC$_c$ scores demonstrated

Table 2. ML Estimations of the *Amborella* Chloroplast Data for Two Competing Trees.

Model[a]	ln L of Tree A	Δln L (Tree B − Tree A)	P* (Tree A)	AIC_c per Site[b]	No. of Model Parameters	F Component Weight[c]
cpREV + F (base model)			0.54	22.39	47	1.0
Base + sel2	−175584.62	−1.68	0.53	22.39	48	1.0
Base + cF4	−175611.95	−1.95	0.45	22.40	51	0.86
Base + cF4 + sel2	−175661.35	2.38	0.45	22.40	52	0.86
Base + cF9	−175685.83	2.16	0.33	22.37	56	0.82
Base + cF9 + sel2	−175419.47	6.85	0.23	22.32	57	0.67
Base + cF20	−175100.47	11.02	0.10	22.28	67	0.66
Base + cF20 + sel2	−174758.52	18.19	0.11	22.28	68	0.65
Poisson + CAT20	−174772.98	18.06	0.14	22.81	66	NA

NOTE.—Tree A, *Amborella* alone at the base of the angiosperm phylogeny; tree B, *Amborella* + water lilies at the base of angiosperms.

[a] sel2 is the selection model with the neutral amino acid profile estimated from equal codon frequency with GARP/FYMINK adjustment.

[b] Corrected AIC score calculated based on tree A. Nearly identical AIC per site scores were obtained if using tree B in the calculation.

[c] Weight for the F component was calculated based on tree A. Similar weight was obtained for each model if using tree B in the calculation. NA, the F component is not applicable for the model.

*P value from the KH test (Kishino and Hasegawa 1989). The null hypothesis is the indicated tree. Small P values suggest strong support for the other tree.

Table 3. ML Estimations of the Breviates Sequence Data for Two Competing Trees.

Model[a]	ln L of the OA Tree	Δln L (OA − BA trees)	P* (BA Tree)	AIC_c per Site[b]	No. of Model Parameters	F Component Weight[c]
LG + F (base model)	−659982.11	−26.29	0.81	30.27	43	1.0
Base + sel2	−659637.36	−28.13	0.82	30.26	44	1.0
Base + cF4	−655589.20	−5.32	0.58	30.07	47	0.57
Base + cF4 + sel2	−655396.84	−5.29	0.58	30.06	48	0.56
Base + cF9	−651183.79	13.72	0.25	29.87	52	0.20
Base + cF9 + sel2	−651158.31	13.92	0.25	29.87	53	0.21
Base + cF20	−647984.93	16.59	0.16	29.72	63	0.18
Base + cF20 + sel2	−647970.47	17.02	0.16	29.72	64	0.18
Poisson + CAT20	−652436.05	0.07	0.50	29.93	62	NA

NOTE.—BA tree, breviates and apusomonads clade tree; OA tree, opisthokonts and apusomonads clade tree.

[a] sel2 is the selection model with the neutral amino acid profile estimated from equal codon frequency with GARP/FYMINK adjustment.

[b] Corrected AIC score calculated based on the BA tree. Identical AIC per site scores were obtained if using OA tree in the calculation.

[c] Weight for the F component was calculated based on the BA tree. Similar weight was obtained for each model if using the OA tree in the calculation. NA, the F component is not applicable for the model.

*P value from the KH test (Kishino and Hasegawa 1989). The null hypothesis is the indicated tree. Small P values suggest strong support for the other tree.

the same trend as in previous examples with cF20 and cF20sel being the best fitting model and the base and Fsel models the worst fitting models. Furthermore, both cF9 and cF20, with or without selection, achieved higher likelihood scores for the OA tree than the BA tree (Δln $L > 13.72$ for the four cases). The CAT20 had a Δln L of only 0.07, and the other models had a negative Δln L. The OA tree was favored, although not in a statistically significant manner according to the KH test, under the cF9, cF20, and their selection models, whereas the other models including CAT20 had less support for the OA tree or supported the BA tree. Furthermore, as in the microsporidia data, the estimated weights of the F component (table 3) for cF9 and cF20 are comparable and less than the corresponding weights for cF4, and the F component weights for the cF models and the corresponding cFsel2 model are also similar.

In the foregoing phylogenomics analyses, the cF models (especially cF20 and cF9), with or without selection, estimated the trees that are more consistent with those that were inferred under advanced phylogenetic methods that employ strategies such as adequate taxa sampling, removing fast-evolving sites and utilizing sophisticated models to handle among-site and among-lineage rate heterogeneities (Brinkmann et al. 2005; Leebens-Mack et al. 2005; Wang et al. 2007; Lartillot et al. 2009; Brown et al. 2013) than the trees estimated under a conventional model employing a single rate matrix. To further evaluate the performance of the selection models on tree estimation, we analyzed the 25 protein data sets simulated under an HB model for site-specific codon frequencies (Holder et al. 2008). Because the data sets were simulated based on the parameters estimated from the mitochondrial cytochrome *b* sequence data, an amino acid rate matrix specific to the mitochondrial proteins (MTrev) (Adachi and Hasegawa 1996) should be a better fitting model than those matrices based on nuclear proteins such as the LG matrix. Indeed, the likelihood of the first

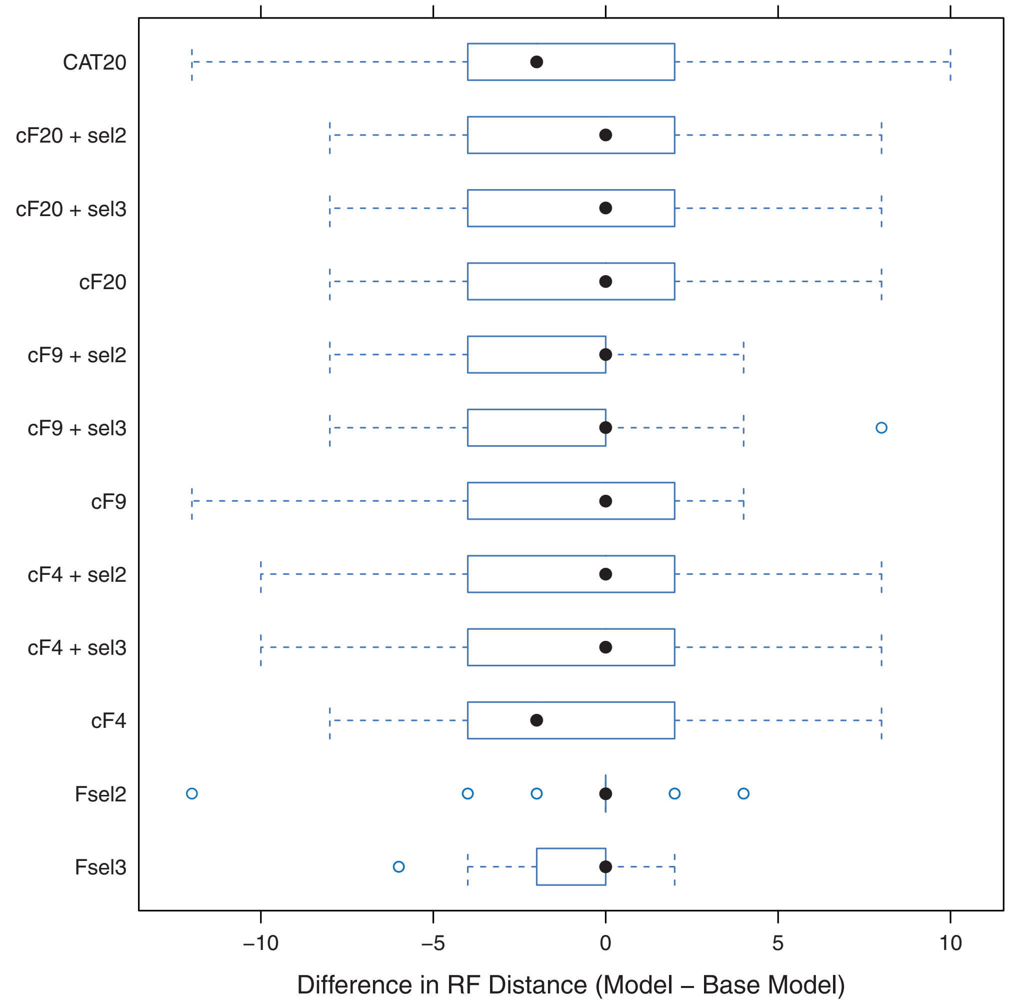

Fig. 3. The difference in the Robinson–Foulds distances between the trees estimated under the models listed on the *y* axis and the base model (MTrev + F + Γ) compared with the generating trees. sel3 is the selection model with the neutral amino acid profile estimated from the 3rd codon positions. sel2 is the selection model with the neutral amino acid profile estimated from equal codon frequency with GARP/FYMINK adjustment.

simulated data set in the “deep tree 1x” data sets (Holder et al. 2008) was much higher under MTrev + F + Γ than under LG + F + Γ (the ln *L*s were −7,532 and −7,704, respectively). Therefore, we used the MTrev as the rate matrix in the base, Fsel, cF, and cFsel models to do tree search for the 25 protein data sets. We used two neutral amino acid profiles for the selection models: the amino acid frequencies expected from equal codon frequency with GARP/FYMINK ratio adjustment (sel2), and the frequencies expected from the nucleotide frequencies at the 3rd codon positions in each data set (sel3). In both cases, to be consistent with the mitochondrial rate matrix (MTrev) used in the models, the vertebrate mitochondrial genetic code (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi, last accessed January 29, 2014) was used to derive the neutral amino acid frequencies. For comparison, the CAT20 model was also applied to the tree estimation for the 25 data sets. Figure 3 shows box plots of the differences in RF distances between the estimated trees and the generating trees for the 12 test models compared with the base model. A paired *t*-test of the mean RF distance under Fsel3 was significantly smaller than the base model ($P = 0.03$), whereas the RF distance under Fsel2 was not significantly smaller than that under the base model ($P = 0.43$). Although the mean RF distance under the cF4, cF4sel2, and cF4sel3 models were all smaller than that under the base model, they were not significant ($P = 0.21$–0.30). Similarly, no significant differences were present for the cF20, cF20sel2, cF20sel3, and CAT20 models relative to the base model. However, the cF9, cF9sel2, and cF9sel3 models gave RF distance differences that were significantly or marginally not significantly ($P = 0.07$, 0.03, and 0.09, respectively) smaller than those of the base model. It is a bit surprising that cF20, cF20sel, or CAT20 models, which have more components of amino acid profiles, did not perform better than the Fsel3, cF9, cF9sel2, and cF9sel3 models. This may be due to the fact that the cF20 profiles were derived from most or all nuclear proteins in the HSSP database, whereas the target data were simulated based on the site patterns of the mammalian mitochondrial proteins, and it is well known that the latter have quite different amino acid compositions than nuclear proteins. The smaller number of the cF9 profiles may be more general to all proteins—nuclear or mitochondrial—which could

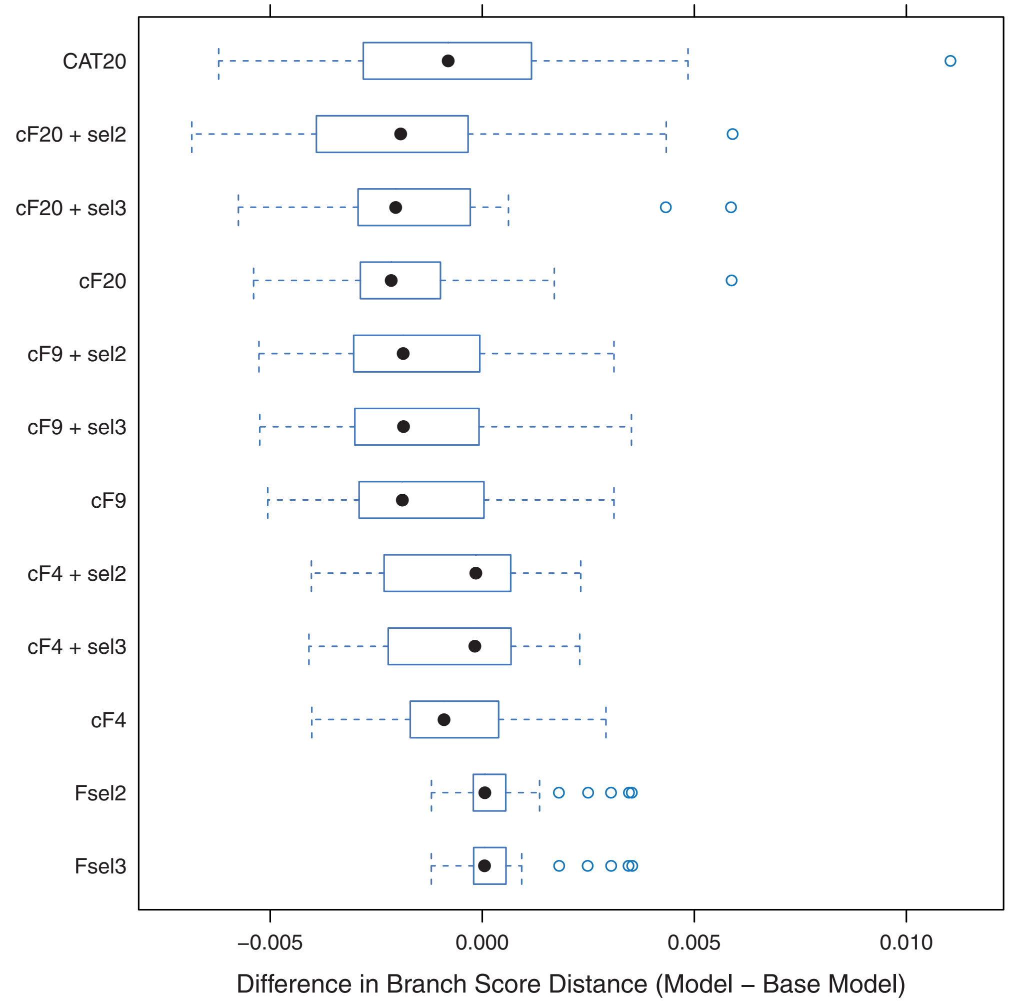

Fig. 4. The differences in the BSDs between the models listed on the *y* axis and the base model (MTrev + F + Γ) compared with the generating trees. See figure 3 legend for labels of the models on the *y* axis.

explain the better tree estimation of the cF9 and cF9sel models.

Figure 4 shows the difference in the branch score distances (BSDs) between the trees estimated under the 12 test models and the base model compared with the generating trees. Again the cF9, cF9sel2, and cF9sel3 had significantly smaller BSD than the base model ($P < 0.003$ in the three cases), followed by cF20, cF20sel2, and cF20sel3 ($P < 0.005$ in the three cases), whereas the other six models (Fsel2, Fsel3, cF4, cF4sel2, cF4sel3, and CAT20) showed no significant differences in BSDs from the base model ($P > 0.07$ in those cases). The results indicated the cF9 and cF20 profiles, with or without selection, tended to estimate more accurate branch lengths than the standard Γ models when the data are simulated under an HB model for the site-specific codon frequencies.

In summary, we have developed a different way of allowing rate matrices of a base model to use different frequencies. Rather than adjusting in the usual way through exchangeabilities, the rate matrices adjust for data set-specific or site-specific frequencies using population genetics principles. We combined the substitution-selection model with an amino acid class frequency mixture model to partially account for site-specific amino acid frequencies. LR tests and AIC$_c$ scores were conducted to show all selection models fit data significantly better than corresponding models without selection for most of the 21 test data sets, and the cF20 with selection models achieved the best likelihoods for most data sets among all the models considered. We further applied the cF and selection models to three difficult phylogenomic problems (the positions of microsporidia and breviate among eukaryotes, and the position of the root in the flowering plant phylogeny) with competing trees. In contrast to conventional models, the cF9, cF20, and their selection models always favored the trees considered more plausible under the most advanced phylogenetic methods and data. Moreover, the cF20 and cF20 selection models showed the highest likelihood gains for the more plausible trees in all three data sets. The KH test is not strictly justified because the trees favored by conventional models are, as such, data dependent. Because it is more difficult to reject an estimated tree than an a priori one, the KH test P values in favor of the alternative trees are likely larger than they would be if a correction for data dependence could easily be applied.

We further used the cF and selection models to estimate trees for data simulated under an HB model for site-specific residue frequencies. Overall, the cF9 and cF9sel models estimated trees closer to the generating trees, in terms of both topology and branch lengths, than the cF20 and cF20sel models, which in turn were better than the base Γ model, cF4, and cF4sel, or CAT20. Because a neutral amino acid frequency profile is important to the model, we explored several means of estimating neutral frequency profiles, including those expected from equal codon frequency with or without the adjustment of the ratio of GARP to FYMINK amino acids (which is an indicator of GC content bias in the underlying genes) and those expected from the nucleotide content at 3rd codon positions. When the coding sequences are known, the latter profile appeared to work better. When only the amino acid sequences are known, the profile based on the equal codon frequency with the adjustment of GARP/FYMINK is preferred over that without the adjustment, although the difference is often not very significant. Other ways of estimating a neutral amino acid profile are possible. For instance, similar to our recent work (Wang et al. 2013), one can simulate sequences under the strictly neutral codon model M0 of Yang et al. (2000) with $\omega = 1$ and with other parameters estimated from the target data. The translated amino acid sequences from the simulated codon sequences may be used as a neutral amino acid profile in the selection models. By properly modeling the site-specific residue fitness in the amino acid substitution process, the cF selection models are valuable for phylogenetic inference and evolutionary studies.

Materials and Methods

An Amino Acid Substitution-Selection Model

Following Halpern and Bruno (1998), a rate matrix for residue changes under selection pressure is defined as follows:

$$Q_{ij}^{(s)} \propto Q_{ij}^{(0)} f_{ij} \tag{1}$$

where $Q_{ij}^{(0)}$ is a rate matrix for the residues under no selection and f_{ij} is the probability of fixation of mutant j given i. According to population genetics theory (Kimura 1962), if the effective population size is N and the relative fitness of j to i is s_{ij}, then the fixation probabilities are:

$$f_{ij} \approx \frac{2s_{ij}}{1 - e^{-2Ns_{ij}}};\, f_{ji} \approx \frac{-2s_{ij}}{1 - e^{2Ns_{ij}}} \tag{2}$$

As in Halpern and Bruno (1998), it follows that $\frac{f_{ij}}{f_{ji}} = e^{2Ns_{ij}} = e^{S_{ij}}$, where $S_{ij} = 2Ns_{ij}$.

Let π_{j_0} and π_j be the stationary frequency of j under neutral evolution and under selection, respectively, then time reversibility requires that

$$\begin{aligned} 1 &= \frac{\pi_i Q_{ij}}{\pi_j Q_{ji}} = \frac{\pi_i}{\pi_j}\frac{Q_{ij}^0 f_{ij}}{Q_{ji}^0 f_{ji}} = \frac{\pi_i}{\pi_j}\frac{\pi_{i_0} Q_{ij}^0}{\pi_{j_0} Q_{ji}^0}\frac{f_{ij}}{f_{ji}}\frac{\pi_{j_0}}{\pi_{i_0}} \\ &= \frac{\pi_i}{\pi_j}\frac{f_{ij}}{f_{ji}}\frac{\pi_{j_0}}{\pi_{i_0}} = \frac{\pi_i}{\pi_j}\frac{\pi_{j_0}}{\pi_{i_0}} e^{S_{ij}} \end{aligned} \tag{3}$$

If we let

$$F_j = \log \frac{\pi_j}{\pi_{j_0}} \tag{4}$$

then equation (3) gives that $S_{ij} = \log \frac{\pi_j}{\pi_{j_0}} - \log \frac{\pi_i}{\pi_{i_0}} = F_j - F_i$. Substituting this into equations (2) and (1), we get:

$$Q_{ij}^{(s)} \propto Q_{ij}^{(0)} \frac{F_j - F_i}{1 - e^{-(F_j - F_i)}} \tag{5}$$

which is equivalent to equation (10) in Halpern and Bruno (1998). Although Halpern and Bruno define this model where i and j are codons, we use it to model evolution at the amino acid level because both can be considered alleles in the population genetics theory, and it is plausible that an amino acid substitution will involve two simultaneous nucleotide changes (Averof et al. 2000; Whelan and Goldman 2004; Doron-Faigenboim and Pupko 2007; Kosiol et al. 2007). Equation (5) applies for standard amino acid rate matrices as well, which allows us to obtain $Q_{ij}^{(0)}$ from these matrices:

$$Q_{ij}^0 = Q_{ij}^{(M)} \frac{1 - e^{-(F_j^{(M)} - F_i^{(M)})}}{F_j^{(M)} - F_i^{(M)}} \tag{6}$$

where $Q^{(M)}$ is any standard amino acid substitution rate matrix such as LG and $F_j^{(M)}$ satisfies equation (4) but with π_j replaced by the equilibrium frequency $\pi_j^{(M)}$ from the rate matrix $Q^{(M)}$. We call this new amino acid model with selection an "F substitution-selection model" (Fsel), as the selection operates on the +F component in the standard protein rate matrices. This can be combined with the standard discretized Γ rate mixture model (Yang 1994) to take into account among-site rate heterogeneity; standard notation would then be Fsel + Γ but for what follows, we leave out the + Γ notation, as rate variation is always included in all the models we discuss.

Estimating Neutral Amino Acid Frequency Profile (π_{j_0})

For the substitution-selection model (eqs. 5 and 6), an amino acid frequency profile under no selection is required to compute the Q matrix under selection. We present three ways of estimating π_{j_0}. The first can be used when the corresponding DNA coding sequences are known for the protein data. Codon frequencies are calculated as the products of 3rd codon nucleotide frequencies, and amino acid frequencies as sums of synonymous codon frequencies. For instance, suppose the frequencies of A, C, G, and T at the third codon position for a gene alignment are a, c, g, and t, respectively. Then, taking the example of Phe because it has two synonymous codons (TTT and TTC), the expected frequency of Phe under neutral evolution is $t \times t \times t + t \times t \times c$ (corrected for the amount of the stop codons). The second way to obtain π_{j_0} is to assume all 61 sense codons have equal frequency (1/61) under no selection; then the amino acid frequencies will be determined by the number of synonymous codons that each amino acid has. This latter approach ignores the fact that amino acid composition is affected by nucleotide bias (the G + C content of the coding sequence). Previous studies have shown that genes with high G + C content encode proteins with higher frequencies of amino acids coded by GC-rich codons (either C or G at the 1st and 2nd codon positions) including G, A, R, and P, whereas genes with high A + T content encode proteins with higher frequencies of amino acids coded by AT-rich codons such as F, Y, M, I, N, and K (e.g., Foster et al. 1997; Wang et al. 2004). Given a protein sequence alignment, the ratio of the frequency of the GARP amino acids to that of the FYMINK amino acids can be determined, and this will correlate with a higher G + C content in the nucleotide coding sequence. The third method for π_{j_0} calculation takes into account the effect of the GC content bias on amino acid composition based on this ratio observed in the protein sequences.

Let A denote the set of amino acids GARP and B denote the set of FYMINK. Under an equal codon frequency model, E, if x is a randomly selected amino acid then x is in A with probability 18/61 and x is in B with probability 12/61, giving the ratio of the GARP to FYMINK amino acids as $\theta_0 = 1.5$. To adjust for a different neutral GARP to FYMINK ratio, θ, we set

$$\pi_{j_0} = \begin{cases} P_E(x = j) \times a & j \in A \\ P_E(x = j) \times b & j \in B \end{cases} \qquad (7)$$

where $b = \frac{2.5}{1+\theta}$ and $a = \frac{\theta \times b}{1.5}$. The frequencies of the other amino acids do not change, except that the resulting amino acid frequencies are rescaled, so that the sum of the frequencies of the 20 amino acids will be 1.0.

An Improved Class Frequency Mixture Model with Selection

Both the standard protein models with the +F variants or the above Fsel models do not adequately account for site-specific substitution processes even when a rates-across-sites process is modeled. A number of "site-heterogeneous" phylogenetic mixture models have been introduced and tested in the last decade, and all appear to better model important site-specific patterns of protein evolution (Lartillot and Philippe 2004; Quang et al. 2008; Wang et al. 2008). One of these was the amino acid site class frequency (cF) mixture model (Wang et al. 2008) that aims to account for the site-specific amino acid substitution patterns for phylogenetic estimation while controlling the number of free parameters used in the model. Under the cF model, amino acid substitutions at a site may be modeled as a mixture of several recurrent site classes each having a specific frequency profile, and the likelihood of a site is a weighted sum of the likelihoods conditional on the amino acid profiles of the site classes. Wang et al. (2008) found this cF mixture model fit data better and gave better phylogenetic estimation than standard models with a single Q matrix, reducing LBA artifacts in simulated data and empirical phylogenomic data. Indeed Lartillot et al. (2009) showed that the cF model is only model other than CAT (Lartillot and Philippe 2004) that can account for substitution saturation for their phylogenomic data.

In the original implementation of the cF model in QmmRAxML (Wang et al. 2008), four site classes were derived from a principal component analysis of the amino acid frequency vectors from the 6,555 aligned positions assembled from 21 conservative protein family data sets. These site classes include a class of predominantly Val, Ile, and Leu; a class of mainly Gly; a class of predominantly Asp and Glu; and a class of a more homogenous amino acid composition. The number of the amino acid sites used in deriving these classes was relatively small and the classification lacked some common site patterns, such as classes that are predominantly RKH, FYW, or AGPST (Dayhoff et al. 1978; Susko and Roger 2007). To get a more representative classification of the known site classes, we utilize two sets of amino acid profiles previously published: 1) the nine component profiles derived from an analysis of the Dirichlet mixture densities over amino acid frequency distributions at the aligned positions of the homologous proteins in the BLOCKS (Henikoff and Henikoff 1991) databases (Sjölander et al. 1996) and 2) the CAT-C20 profiles (Quang et al. 2008) of amino acid site patterns learned from 1,030 protein alignments in the HSSP database (Sander and Schneider 1991). Sjölander et al. (1996) found their nine components mixture of amino acid frequency distributions improved detection of remotely related protein family members. Quang et al. (2008) showed using the CAT-C20 profiles can improve phylogenetic inference especially for the substitution-saturated data. To distinguish the three sets of published cF profiles with the 4, 9, or 20 components, we will refer to them as cF4, cF9, and cF20, respectively. Figure 5 shows the relationship among the three sets of profiles, with cF4 being roughly a subset of cF9 which is in turn a subset of cF20.

The rate matrix for a site having one component in the cF4, cF9, or cF20 profiles is obtained from an empirical rate matrix, Q, using the usual adjustment, via exchangeabilities. The only change is that the role of the data set frequencies is now played by the set of profiles of interest in cF4 (or cF9 and cF20). All models also include an F class, which uses the data set frequencies. So, for instance, cF4 has five classes, one F class for the overall frequencies and four for the cF frequencies. We refer to the model that obtains rate matrices for the F and cF classes, via the usual adjustment to the exchangeabilities, as the cF4 (or, respectively, cF9 and cF20) model. If instead the adjustment incorporates selection for the F component, we refer to it as the cF4sel (or cF9sel and cF20sel) model. Because the class for a given site is unknown, the site likelihood is determined by averaging the partial site likelihoods conditional on each site class profile (Wang et al. 2008):

$$L(x_i) = \frac{1}{g}\sum_{c=1}^{k+1} w_c \sum_{j=1}^{g} P(x_i \mid r_j, \pi_c) \qquad (8)$$

where x_i are data at site i, $K + 1$ are number of cF classes plus the F class ($K = 4, 9,$ and 20 for cF4, cF9, and cF20, respectively), r_j is the rate of a Gamma distribution discretized into one of the g categories ($g = 4$ in all cases studied here) with equal probabilities and w_c's are the weights for the site classes that are estimated by the expectation-maximization algorithm described in Wang et al. (2008).

The cF selection models are expected to partially account for site-specific selection effects on protein evolution. All F

```
cF20 (CAT20)               cF9                        cF4
12 ILM arndcqeghkpstwy     5 ILM arndcqeghkpstwy      1 ILMV arndcqeghkpstwy
 7 AG rdcqehilkmfptwyv     1 ACGST rdqehilkmfwyv      2 G rndcqehilkmfpstwyv
20 DE arcghilkmfpstwyv     7 NDE arcilkmfptwyv        3 DE arncqghilkmfpstwyv
15 NQHY adegilkmptv        3 QK cgilmfpwyv            4 P degilmv
 5 HFWY arndcqegilkmpstv   2 HFWY arndcqegilkmpstv
 4 RQK adcegilmfpstwyv     4 RQK adcegilmfpstwyv
 3 IV arndcqeghkfpstwy     6 IV arndqeghkfpstwy
 8 CHFWY rndqegkp          8 CIMV rndqegkp
16 NDG arcqilkmfptwyv      9 CGPW ailkmfstyv
 1 NST rdqeghilkmfpwyv
 2 ACITV rndqeghlkfpswy
 6 ADE rncghilkmfpwyv
 9 RQLM andcegifpstwyv
10 LMFWY arndcqeghkpstv
11 AQS cilmfpwyv
13 ITV andceghfpswy
14 ACS rndqehilkmfpwyv
17 AP rndceghilkmfwyv
18 QT dcgilmfpwy
19 QEK cgilmfpstwyv
```

Fig. 5. Site patterns (profiles) in the cF4 (Wang et al. 2008), cF9 (Sjölander et al. 1996), and cF20 models. The cF20 model uses the CAT-C20 profiles (Quang et al. 2008). The numbers ahead of each model column indicate the order of the profiles in the original specifications of the model. The amino acids are shown in capital or small letters based on the ratio of the frequencies of the amino acids to the average amino acid frequencies of the LG matrix (Le and Gascuel 2008). Capital letters indicate the ratio greater than 1.5 and small letters indicate the ratio less than 0.85. Amino acids not shown on the list have ratios between 0.85 and 1.5.

and cF substitution-selection models have been implemented in QmmRAxML version 2.0 (http://www.mathstat.dal.ca/~hcwang/QmmRAxML, last accessed January 29, 2014), which requires an additional input file to provide an amino acid frequency profile under no selection (see the above section). Furthermore, to compare with the cF mixture models with or without selection, an empirical profile mixture model (CAT-C20 proposed in Quang et al. 2008) was implemented in QmmRAxML 2.0. The CAT-C20 model (CAT20 for short) uses a proportional rate matrix and the C20 profiles to construct a Q matrix mixture in phylogenetic estimation (Quang et al. 2008). All models used in the article had $+\Gamma_4$ to account for rate heterogeneity across sites.

Data and Model Tests

For the 21-protein family data sets previously assembled in our group (Wang et al. 2008), we first estimated the ML trees under a base model ($\mathrm{LG}+\mathrm{F}+\Gamma$) with four gamma rates. We then estimated the likelihoods of the trees under the Fsel, cF4, cF4sel, cF9, cF9sel, cF20, and cF20sel models. For the selection models, two neutral amino acid profiles (π_{j_0}) were used: in selection 1 (sel1), π_{j_0} was expected from equal codon frequencies and in sel2, π_{j_0} was same as in sel1 but with the GARP/FYMINK ratio adjustment (eq. 7). Because each pair of the sel1 models and corresponding no-selection models (Fsel1 and $\mathrm{LG}+\mathrm{F}+\Gamma$; cF4sel1 and cF4; cF9sel1 and cF9; and cF20sel1 and cF20) have the same number of free parameters, higher likelihood for a model indicate better model fit. The sel2 models have one more free parameter, the ratio of GARP/FYMINK, than the corresponding sel1 models. For comparing the base model with the cF4, cF9, and cF20 models, LR tests with 4, 9, or 20 degrees of freedom were used, as the cF models estimated the weights for the 4, 9, or 20 cF components plus the average amino acid frequency vector of the data set, which sum to 1.0 for each cF model. Because of boundary constraints, usual LR tests are not strictly justified but are expected to give conservative P values (Self and Liang 1987): that is, under the null hypothesis, the probability of a P value less than 0.05 is smaller than 0.05. We further computed for each model, the second order AIC (Akaike 1974) that corrects for sample size (the corrected AIC [$\mathrm{AIC_c}$]), which is calculated as follows:

$$\mathrm{AIC_c} = -2\ln\hat{L} + 2m + \frac{2m(m+1)}{n-m-1}$$

where $\ln\hat{L}$ is the ML estimate under a model, m is the number of the parameters under the model, and n is the number of the aligned positions in the data set. Models with smaller $\mathrm{AIC_c}$ scores are considered to fit the data better.

For comparing the cF4, cF9, and cF20 models, usual LR tests do not apply as the three models were not nested. We, therefore, compared $\mathrm{AIC_c}$ scores for the cF models. In addition, the CAT20 model was applied to estimate the likelihoods of the trees for the 21 data sets. The CAT20 model has 19 free parameters relative to the base model as the estimated weights for the 20 profiles sum to 1.0. Because CAT20 is not nested with the base model or any of the cF models, the $\mathrm{AIC_c}$ scores were used to compare CAT20 with the other models.

To evaluate the performance of the selection models on tree estimation, we analyzed three sets of phylogenomic data. One is the multigene microsporidia data (40 taxa 24,294 sites concatenating from 133 proteins; Brinkmann et al. 2005), which is well known for leading to an LBA artifact in ML analyses with standard protein models where the long-branching microsporidia is placed at the base of eukaryotes close to the long-branching outgroup archaeal species. The gradual removal of fast evolving sites or the use of more complex phylogenetic models that take into account site-to-site heterogeneity (e.g., the CAT model of Lartillot and Philippe [2004]) recovers a tree that groups microsporidia with the fungi (Brinkmann et al. 2005). Another data set of interest was the *Amborella* chloroplast genome data (24 taxa 15,688 sites concatenating from 61 proteins; Leebens-Mack et al. 2005) that has been used to root the tree of angiosperms, which is still a matter of some debate (Goremykin et al. 2013). In this case, two competing trees are *Amborella* alone or *Amborella* + water lilies at the base of the flowering plants and different data types (DNA or protein data) and different phylogenetic models favor one versus the other topology (Leebens-Mack et al. 2005; Wang et al. 2007). A third eukaryotic data set (22 taxa 43,615 sites combined from 159 proteins; Brown et al. 2013) concerns the phylogenetic position of the breviate lineage. Two competing trees regarding the breviate lineage were obtained under the standard LG + F + Γ model with RAxML (Stamatakis et al. 2008) and under CAT-GTR + Γ model with PhyloBayes (Lartillot et al. 2009), respectively (Brown et al. 2013). For the three phylogenomic data sets, we applied a base Γ model with LG or an appropriate rate matrix, cF4, cF9, and cF20, and their selection models to estimate the likelihoods of the predefined competing trees. For comparison, the CAT20 model was also applied to the three data sets. For each data set, we assessed the significance of the difference in log likelihoods for the two competing trees under each of the models using a KH test (Kishino and Hasegawa 1989) to test where there was evidence for or against either of two trees. We also calculated for each data set the corrected AIC$_c$ score for each model. The models having the smallest AIC$_c$ score was the best fitting model.

Finally, we considered data simulated under the HB model of Halpern and Bruno (1998). Holder et al. (2008) analyzed 1,610 mammalian cytochrome *b* sequences to get ML estimates of the site-specific rate parameters under the HB model. They then constructed five trees of 50 taxa selected from the large cytochrome *b* tree and simulated five nucleotide data sets for each tree under the site-specific rates. They found that the standard WAG + F + Γ + I model applied to these data estimated trees with large RF distances (Robinson and Foulds 1981) in the tree topologies to the true generating trees, especially for the 25 data sets with divergent and short sequences (the "deep tree 1x" data: each data set had 50 taxa 1,128 nucleotides, or 376 amino acids). We applied the + F, cF4, cF9, and cF20 models, with or without selection, as well as CAT20, to estimate ML trees for the 25 translated protein data sets. As the nucleotide sequences were available, a neutral amino acid profile based on the nucleotide content at the 3rd codon positions was calculated for each data set and used in the selection models. In addition, we also estimated neutral amino acid profiles expected from equal codon frequency with GARP/FYMINK ratio adjustment for the data sets. As in Holder et al. (2008), we used the generating tree to start the tree search in each case. We used two quantities to compare the estimated trees with the true generating trees: the RF distance and the BSD (Kuhner and Felsenstein 1994). The BSD is the square root of the sum of squared differences in branch lengths. Here, splits that are not present in a given tree are assigned a branch length of 0. The sum is over splits present in either tree. As the generating trees were based on the codon sequences and the branch lengths were estimated in terms of the number of substitutions per codon site, whereas the estimated trees were based on the translated protein sequences and measured as the number of amino acid substitutions per site, the branch lengths were not directly comparable for the estimated trees and the generating trees. We therefore rescaled the branch lengths of both generating trees and the estimated trees by the total tree lengths for each model and for each data set, so that the total tree lengths for each tree after the scaling is 1.0. We used `treedist` in the Phylip package (Felsenstein 2005) to calculate the RF and BSD distances.

Acknowledgments

The authors thank the associate editor and the three reviewers for critical comments on the original manuscript. This work was supported by Discovery grants awarded to A.J.R. and E.S. by the Natural Sciences and Engineering Research Council of Canada (NSERC). A.J.R. acknowledges support from the Canadian Institute for Advanced Research Program in Integrated Microbial Diversity and the Canada Research Chairs program. H.C.W. is currently supported by a CGEB postdoctoral fellowship from the Tula Foundation.

References

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol.* 42:459–468.

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J Mol Evol.* 50:348–358.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automatic Control.* 19:716–723.

Averof M, Rokas A, Wolfe KH, Sharp PM. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287:1283–1286.

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol.* 54:743–757.

Brown MW, Sharpe S, Silberman JD, Heiss A, Lang BF, Simpson AGB, Roger AJ. 2013. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts: implications for the origin of genes involved in multicellularity. *Proc Biol Sci.* 280:1755–1764.

Cao Y, Adachi J, Janke A, Pääbo S, Hasegawa M. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J Mol Evol.* 39:519–527.

Dayhoff MO, Schwartz R, Orcutt BC. 1978. A model for evolutionary change in proteins. Atlas of protein sequence and structure (Vol. 5 suppl. 3). Washington (DC): National Biomedical Research Foundation. Chapter 22, p. 345–352.

Dimmic MW, Mindell DP, Goldstein RA. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac Symp Biocomput.* 5:18–29.

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. *Mol Biol Evol.* 24:388–397.

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

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol.* 44:282–288.

Goldman N, Whelan S. 2002. A novel use of equilibrium frequencies in models of sequence evolution. *Mol Biol Evol.* 19:1821–1831.

Goremykin VV, Nikiforova SV, Biggs PJ, Zhong B, Delange P, Martin W, Woetzel S, Atherton RA, McLenachan PA, Lockhart PJ. 2013. The evolutionary root of flowering plants. *Syst Biol.* 62:50–61.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol.* 15:910–917.

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* 19:6565–657.

Holder MT, Zwickl DJ, Desimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Philos Trans R Soc Lond B Biol Sci.* 363:4012–4013.

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

Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 4:713–719.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J Mol Evol.* 29: 170–179.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol.* 24:1464–1479.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol.* 11:459–468.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

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.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. *Philos Trans R Soc Lond B Biol Sci.* 363:3965–3976.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, depamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. *Mol Biol Evol.* 22: 1948–1963.

Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Robinson DR, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53:131–147.

Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics* 193:557–564.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A.* 107:4629–4634.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9:56–68.

Self S, Liang K. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J Am Stat Assoc.* 82:605–610.

Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci.* 12:327–345.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol.* 57:758–771.

Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol.* 24:2139–2150.

Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190:1101–1115.

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.* 8:331.

Wang HC, Singer GAC, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. *Mol Biol Evol.* 21:90–96.

Wang HC, Spencer M, Susko E, Roger AJ. 2007. Testing for covarion-like evolution in protein sequences. *Mol Biol Evol.* 24:294–305.

Wang HC, Susko E, Roger AJ. 2013. The site-wise log-likelihood score is a good predictor of genes under positive selection. *J Mol Evol.* 76: 280–294.

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.

Whelan S, Goldman N. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* 167:2027–2043.

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, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol.* 25:568–579.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.