General Heterotachy and Distance Method Adjustments

Jihua Wu and Edward Susko

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

Heterotachy is a general term to describe positions in a sequence that evolve at different rates in different lineages. Kolaczkowski and Thornton (2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.) recently described an intriguing heterotachy model that leads to topological bias for likelihood-based methods and parsimony methods. In this article, we show that heterotachy can generally be viewed as multivariate rates-across-sites variation, which can be described as randomly drawing rates (or branch lengths) from a multivariate distribution for each branch at each site. Motivated by this idea, we propose a pairwise alpha heterotachy adjustment model, which gives us much improved topological estimation in the settings by Kolaczkowski and Thornton (2004).

Introduction

Functional constraints on sites in a gene sequence often change through time, causing shifts in site-specific evolutionary rates, a phenomenon called heterotachy (meaning "different speeds") (Fitch 1976; Tuffley and Steel 1997, 1998; Huelsenbeck 2002; Lopez et al. 2002; Kolaczkowski and Thornton 2004; Huelsenbeck et al. 2008; Kim and Sanderson 2008). Tuffley and Steel (1997) were the first to consider the model with site-specific branch lengths, now commonly called the "no-common-mechanism" model. Huelsenbeck et al. (2008) took a Bayesian approach to the no-common-mechanism model and placed independent gamma prior probability distributions on the branch length parameters for each branch. Several special cases of heterotachy models were proposed recently. Covarion models (Fitch and Markowitz 1970; Tuffley and Steel 1998; Galtier 2001; Huelsenbeck 2002) assume that sites have constant probabilities over time of switching between two or more rate categories. Wang et al. (2007) generalized the covarion model to allow evolutionary rates not only to switch between variable and invariable classes but also to switch among different rates when they are in a variable state. A nonstationary model called temporal hidden Markov model was proposed by Whelan (2008), which can distinguish between among-site heterogeneity and among-lineage heterogeneity. Susko et al. (2003) proposed a bivariate model. In this model, they suppose that the rates in two subtrees should be different if it is a heterotachy tree.

Kolaczkowski and Thornton (K&T) (2004) described an interesting heterotachy model where parsimony outperforms misspecified likelihood methods (see also Inagaki et al. 2004). Their research resulted in a lot of discussion (e.g., Philippe et al. 2005; Spencer et al. 2005; Steel 2005), much of which centered on the question of whether maximum parsimony (MP) could be considered better than maximum likelihood (ML) based on their results. Nevertheless, their results did indicate that failing to adjust for heterotachy can lead to incorrect topological estimation. Follow-up work like that of Wang et al. (2008) has supported this conclusion in different settings of heterotachy like that of the covarion model. On the other hand, the work of Wu et al. (2008) indicates that when rates are varying in an independent manner throughout the tree at sites, there will be ways, for instance through distance methods with LogDet distances, of adjusting for heterotachy.

In molecular sequence comparisons, rate and time are intrinsically confounded (Felsenstein 1981; Thorne and Kishino 2002; Yang and Yoder 2003). The likelihood of the data depends on the distances or branch lengths (the product or integral of rates and times) but not on evolutionary rates and times individually.

We will give a general definition of heterotachy as multivariate rates-across-sites (RAS) variation, where the rates or branch lengths are modeled by an arbitrary distribution. We also show that all the commonly considered heterotachy models can be considered special cases of this general definition. We use this result to establish that, for pairs of taxa, heterotachy is an usual, univariate RAS variation. Motivated by this characterization of heterotachy, we assume that rates between two taxa follow different distributions. For convenience, gamma distributions (Yang 1993, 1994; Tuffley and Steel 1998) with different shape parameters for different pairs are used in this article. Through ML estimation, we find the best distance and the best (shape) parameter for the corresponding distribution for each pair of taxa (see fig. 1). Then, the Neighbor-Joining (NJ) method (Saitou and Nei 1987) is used to find the best tree topology.

Similarly as in Kolaczkowski and Thornton (2004), we simulated replicate DNA sequence alignments with two symmetrical rate partitions along a four-taxon tree. Our pairwise alpha heterotachy adjustment (PAHA) consistently showed better phylogenetic accuracy (the fraction of replicates from which the true tree was recovered) on simulated data than either uncorrected ML or MP methods.

Methods

Heterotachy

Although there are a variety of heterotachy models, they all turn out to be related. In this section, we provide two equivalent general definitions for heterotachy. Then we show that the equal rates (ER) model, RAS model (Yang 1993), K&T four-taxon model (Kolaczkowski and Thornton 2004), bivariate model (Susko et al. 2003), random-effect rates variation (RERV) model (Wu et al. 2008), and covarion model (Tuffley and Steel 1998) are all special cases of heterotachy. Earlier description of the term "heterotachy" (Fitch 1976) did not exclude a possible

Key words: heterotachy, covarion model, distance method, tree estimation.

E-mail: jihuaw@mathstat.dal.ca.

Mol. Biol. Evol. 26(12):2689–2697. 2009
doi:10.1093/molbev/msp184
Advance Access publication August 17, 2009

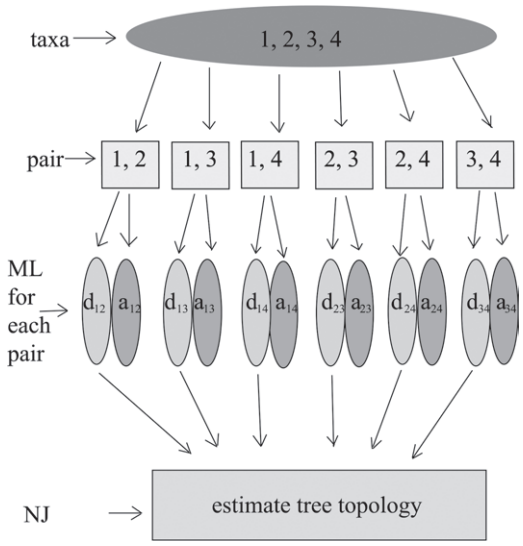

FIG. 1.—The pairwise alpha heterotachy adjustment.

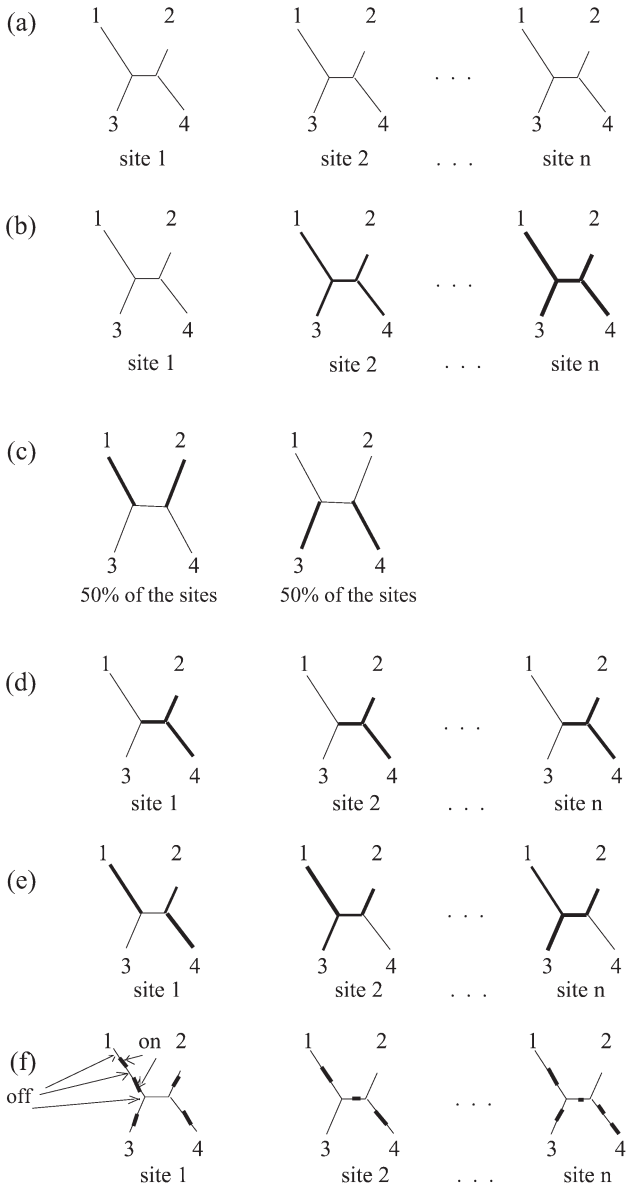

FIG. 2.—The distribution of RAS and lineages under six models of evolution. Each tree plot describes the distribution of rates across lineages for a particular site under the considered model. Rates are represented by different line thicknesses. (*a*) ER model; (*b*) RAS model; (*c*) K&T four-taxon model (settings); (*d*) bivariate model; (*e*) RERV model; (*f*) covarion model.

dependence across sites, which is more general than what we define below. However, without the assumption of independence, the model implementation becomes much more difficult for a number of reasons including the difficulty of specifying the nature of dependence.

Definition 1: Heterotachy as multivariate RAS variation.

Generally, heterotachy can be viewed as independent draws of rates from some multivariate distribution G for each site. We let $r_i^{(h)}$ denote the average rate at site h for edge i. For an unrooted tree with m taxa, there are $2m-3$ edges. Denote a rate vector $\boldsymbol{r}^{(h)} = (r_1^{(h)}, r_2^{(h)}, \ldots, r_{2m-3}^{(h)})$, then with n sites, $\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \ldots, \boldsymbol{r}^{(n)}$ are independent draws from some multivariate distribution G, where $E_G[r_i^{(h)}] = 1$.

Definition 2: Heterotachy as multivariate edge length variation across sites and lineages.

In this model description, we draw vectors of edge lengths $\boldsymbol{t}^{(h)} = (t_1^{(h)}, t_2^{(h)}, \ldots, t_{2m-3}^{(h)})$ in an independent and identically distributed fashion from some distribution H. We can then define overall, average edge lengths as

$$t_i = E_G[t_i^{(h)}], \quad i = 1, 2, \ldots, 2m-3. \tag{1}$$

The two definitions are equivalent. Definition 1 can be obtained through the following equation:

$$r_i^{(h)} = t_i^{(h)} / t_i. \tag{2}$$

We will show that all common heterotachy models are a special case of the general heterotachy model (either by Definition 1 or by Definition 2).

(a) ER model (fig. 2*a*). Because the interpretation of edge lengths is the expected number of substitutions, the common rate is 1. If we set $r_j^{(h)} = 1$, we get an ER model.

(b) RAS model (fig. 2*b*). In this model, a single rate $r^{(h)}$ is drawn independently at each site and acts as an overall rate multiplier for the entire tree. If we set $r_j^{(h)} = r^{(h)}$, we get the RAS model (Yang 1993, 1994).

(c) K&T four-taxon model (figs. 2*c* and 3). In this model (Kolaczkowski and Thornton 2004), a proportion of sites was drawn from the following tree, in Newick format,

$$((A:0.75, B:0.05):0.1, (C:0.75, D:0.05)),$$

and another proportion of the sites is drawn from another tree

$$((A:0.05, B:0.75):0.1, (C:0.05, D:0.75)).$$

If the proportions drawn were random, this would correspond to Definition 2, so the K&T four-taxon model is a special case of the general heterotachy model.

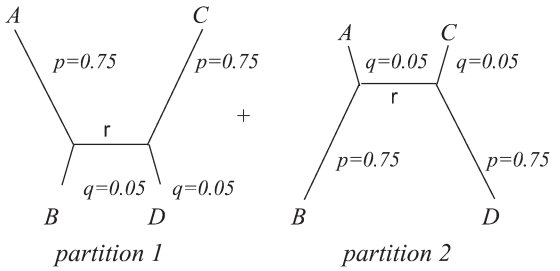

FIG. 3.—Simulation settings for the K&T-type scenarios. A certain proportion of sites are generated from the tree in partition 1 and the rest from partition 2.

(d) Bivariate model (fig. 2*d*). We split the tree into two subtrees and for each site h, we let the rates at one subtree equals $r_a^{(h)}$ and the other subtree equals $r_b^{(h)}$. Without loss of generality, assume that the edges $1,2,\ldots,m_1$ are in the first subtree and $m_1+1,m_2+1,\ldots,2m-3$ are in the second. Then setting $r_i^{(h)}=r_a^{(h)}$, $i=1,2,\ldots,m_1$, and $r_i^{(h)}=r_b^{(h)}$, $i=m_1+1,m_2+1,\ldots,2m-3$, we see that the bivariate model (Susko et al. 2003) is a special case of Definition 1.

(e) RERV model (fig. 2*e*). If for each branch j, we independently randomly draw $r_j^{(h)}$, from a distribution, for example, $\Gamma(\alpha_j)$, we get the RERV model (Wu et al. 2008).

(f) Covarion model (fig. 2*f*). The covarion model seems different than the heterotachy model described here because rates can vary along a fixed edge. For a pair of sequences, the covarion model gives results identical with those of a suitably chosen RAS model (Tuffley and Steel 1998). This formulation applies as well to a pair of sequences at end nodes of a branch in a tree. Thus, for each site and branch, the covarion process is equivalent to assignment of a random rate. The stochastic process that defines a covarion model gives rise to a complicated dependence between these rates. Nevertheless, Definition 1 applies.

Heterotachy as Differing RAS Distributions for Pairs

From the above discussion, all common heterotachy models satisfy Definitions 1 and 2. However, it is difficult to determine which of the above models gives the most realistic model of heterotachy. The difficulty is in specifying the form of dependence of rates across lineages. We show here that this problem can be avoided by considering distances. For pairs of taxa, heterotachy turns out to be simply RAS variation with different RAS distributions for different pairs.

We start by considering a multivariate discrete distribution for the edge lengths. This is the case in which there are a finite number of edge lengths that might randomly be selected at a site. We assume that there are g groups of edge lengths and that the jth group will be selected with probability w_j.

Now, consider a pair of taxa (fig. 4). In order to differentiate the rates and edge lengths in different rate groups from those at different sites, we use upper case for groups. Then, for group i, the distance between taxa 1 and 4 is $D_{14}^{(i)}=T_1^{(i)}+T_5^{(i)}+T_4^{(i)}$, $i=1,2,\ldots,g$. Thus, for this pair of taxa, at each site h, drawing edge lengths at random is

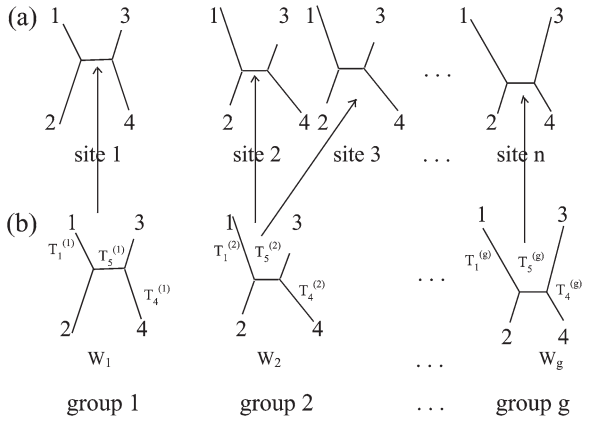

FIG. 4.—(*a*) For each site, characters are generated from some group in (*b*), which makes the edge lengths different at different sites; (*b*) there are g groups of partitions, each partition comes from an ER model, and group i has weight w_i.

the same as drawing distances at random where with probability w_i, $d_{14}^{(h)}=D_{14}^{(i)}$, $i=1,2,\ldots,g$.

Alternatively, we can view this as an RAS model as follows. Let

$$d_{14}=\sum_{i=1}^{g} w_i D_{14}^{(i)} \tag{3}$$

and

$$R_{14}^{(i)}=D_{14}^{(i)}/d_{14},\quad i=1,2,\ldots,g. \tag{4}$$

Then the model is the same as the one in which taxa 1 and 4 have distance d_{14} and rate $r_{14}^{(h)}$, where the rate at each site h is selected at random and with probability w_i, $r_{14}^{(h)}=R_{14}^{(i)}$. Here $\sum w_i R_{14}^{(i)}=1$, so that this is a conventional RAS model.

Up to now, we have assumed that the multivariate distribution of edge lengths describing heterotachy is discrete. Suppose now that it is continuous. Again, considering taxa 1 and 4 for illustration, the random distance at a site h is

$$d_{14}^{(h)}=t_1^{(h)}+t_5^{(h)}+t_4^{(h)}.$$

But now, instead of having $d_{14}^{(h)}$ drawn from a discrete distribution, it is drawn from a continuous density, $h_{14}(x)$; this density can, in principle, be determined from the multivariate distribution for $t_1^{(h)}$, $t_5^{(h)}$, and $t_4^{(h)}$. Alternatively, we can view this as an RAS model in a similar way as we did in the discrete case by letting

$$d_{14}=\int x h_{14}(x)\mathrm{d}x.$$

Defining random rates at sites as $r_{14}^{(h)}=d_{14}^{(h)}/d_{14}$, the RAS density is

$$g_{14}(r)=d_{14}\cdot h_{14}(d_{14}r)$$

and satisfies the usual constraint that $\int r\cdot g_{14}(r)\mathrm{d}r=1$.

Using Distance Methods to Adjust for Heterotachy

From the above discussion, heterotachy can be modeled as an RAS variation for sites. In our implementation, for a pair of taxa, we assume that the rate at each site is drawn from a distribution $R(r,\alpha)$; we use a gamma

distribution (Yang 1993, 1994; Tuffley and Steel 1998) as a parametric model. We obtain parameter estimates (the distances and the shape parameter α) through ML estimation. Under heterotachy, different distributions apply to different pairs. For each pair of taxa, we find the best distance and the best shape parameter α (see fig. 1). After this, the NJ method (Saitou and Nei 1987) is used to find the best tree topology.

Suppose that for a pair of taxa of interest, we have character states (a_i, b_i) at the ith site in sequences A and B. Let $P_{a_i b_i}(r^{(i)}d)$ denote the probability that character state a_i is substituted by b_i along a distance of length d and a evolutionary rate $r^{(i)}$. Then the conditional probability of (a_i, b_i) at the site, given rate $r^{(i)}$ and distance d, is

$$P(a_i, b_i | r^{(i)}, d) = \pi(a_i) P_{a_i b_i}(r^{(i)} d), \tag{5}$$

where d and $r^{(i)}$ are the distance and evolutionary rate from one taxa to the other at the site and $\pi(a_i)$ is the probability of character state a_i. Because $r^{(i)}$, $i = 1, 2, \ldots, n$, is unknown, we calculate the unconditional probability of the observed data, (a_i, b_i), as

$$P(a_i, b_i; d, \alpha) = \int_0^{+\infty} \pi(a_i) P_{a_i b_i}(rd) R(r, \alpha) \mathrm{d}r.$$

The likelihood is obtained by multiplying the probabilities over all sites, which is

$$L(\alpha, d) = \prod_{i=1}^{n} \int_0^{+\infty} \pi(a_i) P_{a_i b_i}(rd) R(r, \alpha) \mathrm{d}r. \tag{6}$$

Let n_c denote the number of character states; four for nucleotide models and 20 for amino acid models. The total number of combinations for pairs of character states is n_c^2. For example, for nucleotide data, there are four character states: A, C, G, and T, so the combinations for pair of character states are (A,A), (A,C), …, (T,T), 16 in total ($4^2 = 16$). In order to save computation time, we count the frequency for each character state combination as $n_{ij}, i, j = 1, \ldots, n_c$, and reorganize equation (6) as

$$L(\alpha, d) = \prod_{i=1}^{n_c} \prod_{j=1}^{n_c} \left\{ \int_0^{+\infty} \left[\pi(c_i) P_{c_i c_j}(rd) R(r, \alpha) \right] dr \right\}^{n_{ij}}, \tag{7}$$

where $\pi(c_i)$ is the probability of character state c_i and $P_{c_i c_j}(rd)$ is the transition probability from character state c_i to c_j with distance d and rate r. Maximizing the log likelihood in equation (7) gives estimates of d and α. After we estimate d for each pair of taxa, we can use the NJ method to find the best tree topology.

For the PAHA method, there is a question about whether the parameters α and d are identifiable or not; α and d are unidentifiable if two different sets of α and d values give exactly the same probability of observing any given pair of sequences. The model may include additional (rate matrix) parameters, such as the transition/transversion ratio κ for the HKY85 model (Hasegawa et al. 1985) and κ_1 and κ_2 for the TN93 model (Tamura and Nei 1993). We treat these as known and constant across pairs. In practice, they can be estimated by software including Tree-Puzzle (Schmidt et al. 2002). For amino acid data, empirical rate matrices, such as the Jones–Taylor–Thornton (JTT) rate matrix (Jones et al. 1992), with no unknown parameters are commonly used.

According to Wu et al. (2008), if the rate matrix used in the PAHA method has more than two distinct eigenvalues, α and d will be identifiable. Because the rate matrix used in this paper is the JTT rate matrix, which has 19 distinct eigenvalues, the parameters are identifiable at least for amino acid data. For nucleotide data, the situation is different. For example, if the JC69 (Jukes and Cantor 1969) or F81 (Felsenstein 1981) rate matrix is used in the PAHA method, the parameters are unidentifiable, but if the HKY85, TN93, or general time reversible (GTR) rate matrix (Tavaré 1986) is used, the parameters are identifiable (assuming that these three models do not degenerate to the JC69 or F81 model). The difference here is that there are just two distinct eigenvalues for the JC69 rate matrix, but there are at least three distinct eigenvalues for the HKY85, TN93, and GTR rate matrix.

Results

Simulations

Kolaczkowski and Thornton (2004) simulated data sets under each set of conditions using the JC69 (DNA) model. However, the parameters α and d under JC69 model are unidentifiable. The JTT rate matrix has 19 distinct eigenvalues, so the parameters are identifiable for amino acid data (Wu et al. 2008). For these reasons, we simulated and analyzed amino acid data under JTT model (C code implementing PAHA estimation is available on request).

Similarly as in Kolaczkowski and Thornton (2004), we simulated amino acid sequences along a four-taxon tree ((A,B),(C,D)) with two independent partitions that were concatenated into one heterogeneous alignment (see fig. 3). In partition 1, long terminal branches equal $p = 0.75$ and lead to A and C and short terminals equal $q = 0.05$ and lead to B and D. In partition 2, terminal branches to B and D have length p, whereas A and C have length q. The internal branch length $r \in (0.0, 0.5)$ is equal in both partitions. The two partitions were of equal size unless otherwise noted. One thousand replicate alignments of 1,000 and 10,000 characters were simulated under each set of conditions using the JTT model through seq-gen (Rambaut and Grassly 1997). We simulated each of the two partition of sites (trees) with the ER model, because the two trees have different branch length, and then mixed them together to get a general heterotachy tree.

We also simulated four taxon amino acid sequences through seq-gen-aminocov (Ané et al. 2005). We set the model of Tuffley and Steel with ON frequency 0.25 and speed of covarion evolution 1.75. The evolution speed is the average number of switches per (average) substitution. One thousand replicate alignments of 1,000 and 10,000 characters were simulated under the JTT model.

After getting the simulated sequences, we compared percentages of correctly recovery tree among the PAHA method, uncorrected ML method, and MP method. Here, uncorrected ML method means that we estimated tree topology through ML method but used different model as that of in the simulations. We use ML + ER and ML + RAS to denote the method of estimating the trees always using

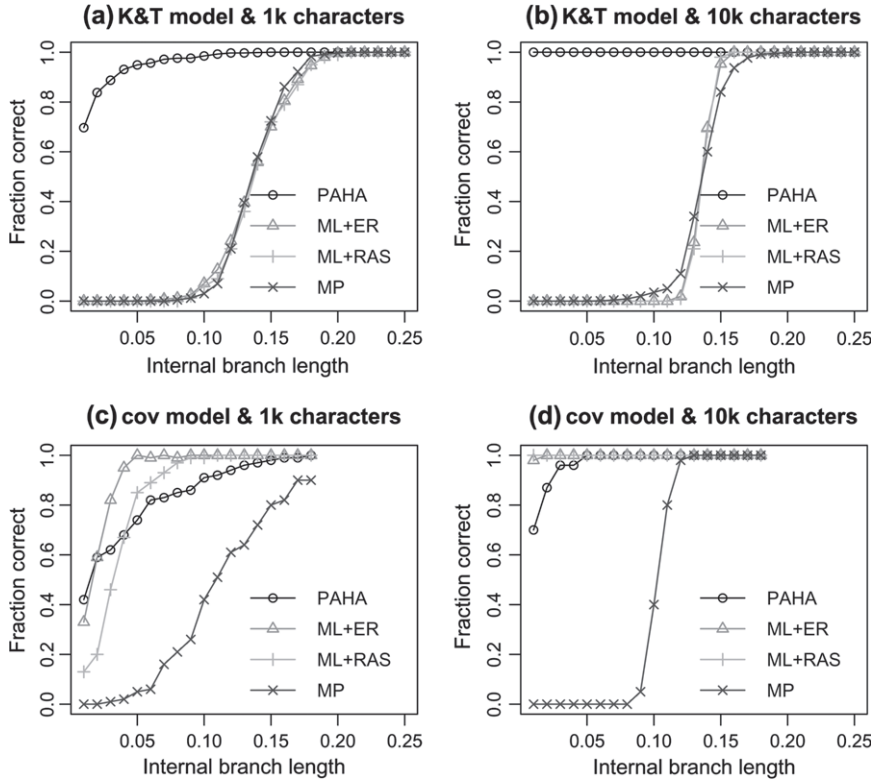

FIG. 5.—The proportions of correct reconstructions as a function of internal edge length for PAHA, uncorrected ML, and MP methods when data are generated according to the K&T scenario of figure 3.

ML method but assuming that ER and RAS (a single shape parameter alpha is estimated) models were applied respectively.

Figure 5*a* and *b* gives the proportion of correct reconstructions as a function of the internal branch length under the K&T four-taxon model simulations. Under the condition of strong heterotachy considered here ($p = 0.75,\ q = 0.05$), the accuracy of PAHA is much better compared with ML + ER, ML + RAS, and MP.

Figure 5*c* and *d* gives the proportion of correct reconstructions as a function of the internal branch length under the covarion model simulations. We can see that the accuracy of PAHA is not apparently worse than ML + ER, ML + RAS, and MP. However, we are a little surprise that the accuracy of ML + RAS is worse than that of ML + ER in 1,000 characters case (fig. 5*c*).

The proportion of sites varies from 0.0 to 1.0 in the first partition, but the internal branch lengths and the number of characters are fixed in each case. From figure 6, we can see that heterotachy does not affect the accuracy except when the proportion of sites in partition 1 is close to 0.1 or 0.9. Compared with ML + ER, ML + RAS, and MP, we can see that PAHA is much more resistant to heterotachy. Also the accuracy of the PAHA method is much better compared with ML + ER, ML + RAS, and MP, especially when the number of sites is large.

Chloroplast Data

The chloroplast data consist of 61 concatenated protein-coding genes and have 15,688 sites for 24 taxa. These data have been considered before, and the inferred relationship between Amborella and the Nymphaeales varied when using different methods of phylogenetic reconstruction, models of molecular evolution, and subsets of taxa (Barkman et al. 2000; Graham and Olmstead 2000; Zanis et al. 2002; Stefanović et al. 2004; Susko and Roger 2007). The branching order of Amborella and the Nymphaeales relative to each other and the rest of the angiosperms remains uncertain. Using the chloroplast genome data, Leebens-Mack et al. (2005) found weak support for Amborella and the Nymphaeales at the base of the angiosperms.

The original sequences were used to generate 100 data sets according to the nonparametric bootstrap scheme of Felsenstein (1985). These data sets were then used to estimate 100 distance matrices for the ER and RAS models and PAHA. We use NJ + ER, NJ + RAS, and NJ + PAHA to denote the method of distance matrix estimation; tree estimation is always through the NJ method. For the NJ + RAS method, a single shape parameter alpha is estimated when we estimate the distance matrix. For each method, bootstrap support values were calculated resulting in the trees labeled NJ + ER, NJ + RAS, and NJ+PAHA in figure 7. Tree-Puzzle (Schmidt et al. 2002) was used to estimate distance matrices for ER and RAS. The programs Seqboot, Neighbor, and Consense in the PHYLIP package (Felsenstein 1993) were used for bootstrap generation, application of the NJ algorithm, and bootstrap summary.

There are several differences between the trees in figure 7 as well as those of Leebens-Mack et al. (2005), which is based on nucleotide data limited to the first two

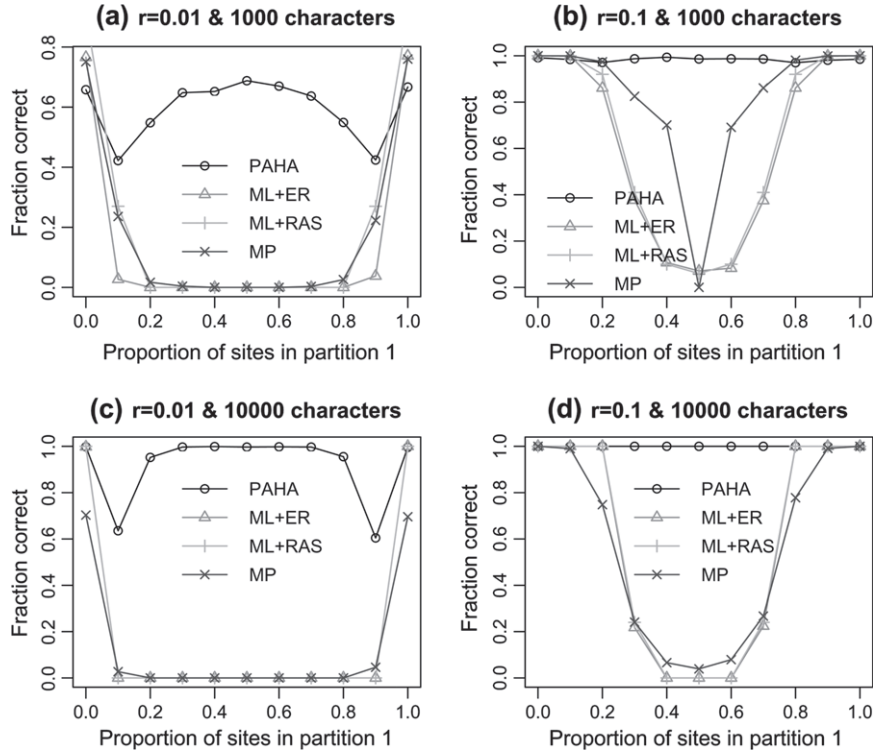

FIG. 6.—The proportions of correct reconstructions as a function of proportion of sites in partition 1 for PAHA, uncorrected ML, and MP methods when data are generated according to the K&T scenario of figure 3.

codon positions in contrast to the amino acid analysis here. The first that we comment on is the placement of the outgroup taxa *Marchantia* with *Psilotum* with 100% bootstrap support instead of with *Physcomitrella* as occurred with 100% bootstrap support in the NJ + ER tree and the trees of Leebens-Mack et al. (2005) in their figure 4A–C. Notably, the *Marchantia* with *Psilotum* grouping of NJ + PAHA occurred as well in the NJ + RAS tree, the other method that adjusts for rate variation, albeit with lower (57%) bootstrap support. The difference in bootstrap support indicates that RAS variation is being inferred for the distances between these taxa and some of the others in the tree. For each pair of taxa coming from the three taxa *Marchantia*, *Psilotum*, and *Physcomitrella*, the estimated α was much smaller than the α calculated for other pairs of taxa, which suggests that heterotachy may be present in this portion of the tree. Therefore, the *Marchantia* with *Psilotum* grouping might be correct and estimated because of adjustment for heterotachy through the PAHA method.

Other differences concerned the relative placements of *Typha*, *Yucca*, and *Calycanthus*. For the NJ + PAHA tree, these were the same as for the ML tree of Leebens-Mack et al. (2005). The placement of *Calycanthus* in the NJ + PAHA tree was more uncertain, however, with 58% bootstrap support. The placement of *Typha* and *Yucca* is the same as in the NJ + RAS tree and differs from the NJ + ER tree. The placement of *Typha* and *Yucca* in the NJ + ER tree is likely due to a failure to adjust for rate variation of any sort.

The final difference worth noting is the NJ + PAHA placement of Amborella and the Nymphaeales, consistent with Leebens-Mack et al., at the base of the angiosperms with relatively large bootstrap support. This placement has been the source of considerable debate (Goremykin et al. 2003, 2004; Soltis et al. 2004; Lockhart and Penny 2005; Martin et al. 2005; Jansen et al. 2007; Moore et al. 2007). Interestingly, the NJ + RAS and NJ + ER trees place this group with the monocots. It is possible that (lack of) adjustments for what is, in actuality, heterotachy are what is causing the differences in the estimated trees.

Discussion

A large number of different heterotachy models have been proposed during the last 15 years (Yang 1993; Tuffley and Steel 1998; Susko et al. 2003; Kolaczkowski and Thornton 2004; Wu et al. 2008). These heterotachy models are seemingly very different. Some are "punctuated" models, like the K&T four-taxon model, some are "gradual" models, like the RAS model. It will be valuable for future work to recognize that there is a framework (the general heterotachy model) for all these models. When considered from the perspective of pairs, it is also useful to note that heterotachy is an RAS variation but with different RAS distributions for different pairs. Based on this observation, we proposed the PAHA method that can be used to estimate the tree topology.

An alternative method to estimate the tree topology is full ML estimation, which needs a tree and a model of heterotachy to specify probabilities of patterns for all taxa. The ML estimate of the phylogeny is the tree that maximizes the probability of the sequence data. Difficulties here

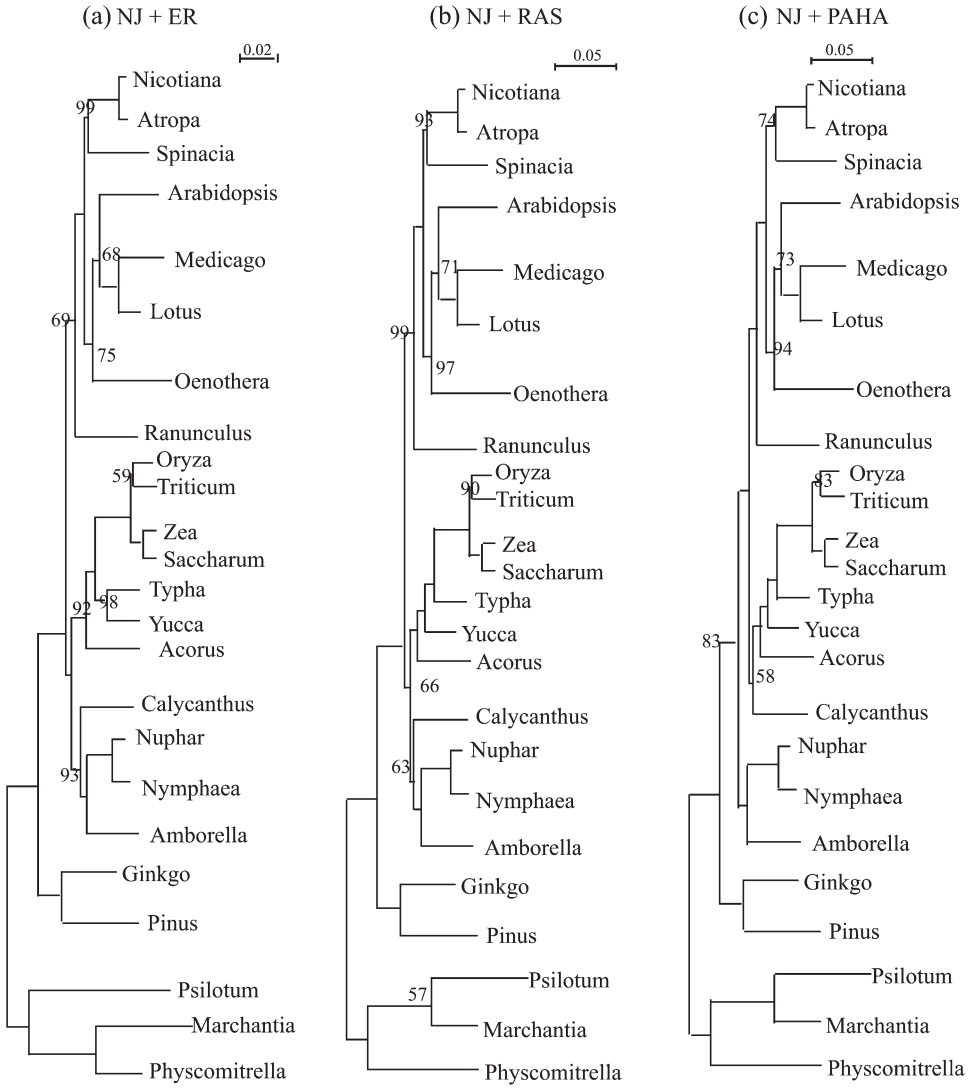

FIG. 7.—(*a*) ER model + NJ, (*b*) RAS model + NJ, and (*c*) PAHA method + NJ. Phylogenies for the chloroplast data estimated using distances were obtained assuming a JTT substitution model. Most nodes on each phylogeny were recovered in 100% of the bootstrap replicates, and only values <100% are shown for each node. NJ + PAHA place Amborella and the water lilies as basal lineages in the angiosperm phylogeny with 100% support values.

include that the pruning algorithm required for estimating the edge lengths is time consuming for large numbers of taxa. With tree searching, computation is even more expensive. The more serious problem is that full ML estimation requires a model of heterotachy that includes specification of the dependence structure in the multivariate distribution. Because there are many dependence structures, for example, the RAS model, the K&T four-taxon model, or the covarion model, which heterotachy model should be used in full ML estimation is uncertain. Compared with full ML, PAHA does not depend on the tree structure and avoids the difficulties of modeling dependence structures because PAHA is based on pairs of sequences. For the same reason, PAHA also has computational advantages.

There is a concern that for PAHA, there may be not enough information in pairs of sequences to estimate an α and a distance. Our results suggest that there is enough information for large data sets with at least 1,000 sites, although we expect that full ML would show performance improvements if heterotachy is correctly modeled. From the results of our simulations, we found that the performance will be improved with larger sequence length. We expect that the PAHA method will be more useful for large concatenated sequences. The PAHA method performs well under the K&T four-taxon model and has much better accuracy than the uncorrected full ML and MP methods for phylogeny reconstruction in most cases (see figs. 5 and 6) and is also simple to implement. The simulations and the analysis of chloroplast data indicate that the PAHA method performed well.

For phylogeny reconstruction, Ninio et al. (2007) proposed five distance-based methods, including a method where pairwise α's were estimated. Their interest was in estimation under an RAS model rather than as a means of correcting for heterotachy. They found that a method they refer to as the iterative posterior method achieved better results than the pairwise α method. However, their iterative posterior method had one shape parameter (α) for

the whole tree, so that the iterative posterior method cannot deal with general heterotachy. The pairwise α method estimates the α parameter for each pair of sequences separately. These α's contain the information about heterotachy. Even though sequence generation used a single α for the entire tree, for many generated α's in figure 2 in Ninio et al. (2007), the performance of the pairwise α method was not much worse than the global α estimation methods. A notable exception occurred with $\alpha > 1$. If $\alpha > 1$, the rate variation process becomes more like an ER model. We suggest caution when PAHA gives a lot of α's larger than one and comparison with the ER tree. The reason for bias in case of $\alpha > 1$ is unclear but deserves further investigation.

Finally, we would note that there has been some work on testing for heterotachy (Lockhart et al. 1998; Ané et al. 2005; Lockhart and Steel 2005; Baele et al. 2006; Gruenheit et al. 2008). Our emphasis here has been on estimation in the presence of heterotachy. However, the differences and similarities of the α's estimated through a pairwise approach may also be used for testing, which will be the focus of future work.

Acknowledgments

We wish to thank two anonymous reviewers for helpful comments. This research was supported by a Discovery Grant awarded to E.S. by the Natural Sciences and Engineering Research Council of Canada.

Literature Cited

Ané C, Burleigh JG, McMahon MM, Sanderson MJ. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol. 22(4):914–924.

Baele G, Raes J, Van de Peer Y, Vansteelandt S. 2006. An improved statistical method for detecting heterotachy in nucleotide sequences. Mol Biol Evol. 23:1397–1405.

Barkman TJ, Chenery G, McNeal JR, Lyons-Weiler J, Ellisens WJ, Moore G, Wolfe AD, DePamphilis CW. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc Natl Acad Sci USA. 97:13166–13171.

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.

Felsenstein J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author. Seattle (WA): Department of Genetics, University of Washington.

Fitch WM. 1976. The molecular evolution of cytochrome c in eukaryotes. J Mol Evol. 8:13–40.

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet. 4:579–593.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 18:866–873.

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2003. Analysis of the Amborella trichopoda chloroplast genome sequence suggests that Amborella is not a basal angiosperm. Mol Biol Evol. 20:1499–1505.

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 21:1445–1454.

Graham SW, Olmstead RG. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am J Bot. 87:1712–1730.

Gruenheit N, Lockhart PJ, Steel M, Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol. 25:1512–1520.

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.

Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol. 19:698–707.

Huelsenbeck JP, Ané C, Larget B, Ronquist F. 2008. A Bayesian perspective on a non-parsimonious parsimony model. Syst Biol. 57(3):406–419.

Inagaki Y, Susko E, Fast NM, Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1 alpha phylogenies. Mol Biol Evol. 21:1340–1349.

Jansen RK, Cai Z, Raubeson LA, et al. (16 co-authors). 2007. Analysis of 81 genes from 64 chloroplast genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA. 104:19369–19374.

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

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian Protein Metabolism. New York: Academic Press, p. 21–132.

Kim J, Sanderson M. 2008. Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. Syst Biol. 57(5):665–674.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Leebens-Mack J, Raubeson LA, Cui LY, 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.

Lockhart P, Penny D. 2005. The place of Amborella within the radiation of angiosperms. Trends Plant Sci. 10:201–202.

Lockhart PJ, Steel MA. 2005. A tale of two processes. Syst Biol. 54:948–951.

Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol. 15:1183–1188.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19:1–7.

Martin W, Deusch O, Stawski N, Grünhelt N, Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:1360–1385.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genomic-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA. 104:19363–19368.

Ninio M, Privman E, Pupko T, Friedman N. 2007. Phylogeny reconstruction: increasing the accuracy of pairwise distance estimation using Bayesian inference of evolutionary rates. Bioinformatics 23(2):e136–e141.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. Evol Biol. 5(1):50.

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.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing evolutionary trees. Mol Biol Evol. 4:406–425.
Schmidt HA, Strimmer K, Vingron M, Haeseler AV. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.
Soltis D, Albert V, Savolainen V, et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and ending incongruence: a cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.
Spencer M, Susko E, Roger AJ. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol. 22:1161–1164.
Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'? Trends Genet. 21:307–309.
Stefanović S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol. 4:35.
Susko E, Field C, Blouin C, Roger AJ. 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. Syst Biol. 52:594–603.
Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol. 24:2139–2150.
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.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences. Am Math Soc. 17:57–86.
Thorne JL, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst Biol. 51(5): 689–702.
Tuffley C, Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol. 59(3):581–607.
Tuffley C, Steel MA. 1998. Modelling the covarion hypothesis of nucleotide substitution. Math Biosci. 147:63–91.
Wang HC, Spencer M, Susko E, Roger AJ. 2007. Testing for covarion-like evolution in protein sequences. Mol Biol Evol. 24(1):294–305.
Wang HC, Susko E, Spencer M, Roger AJ. 2008. Topological estimation biases with covarion evolution. J Mol Evol. 66: 50–60.
Whelan S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol Biol Evol. 25:1683–1694.
Wu J, Susko E, Roger AJ. 2008. An independent heterotachy model and its implications for phylogeny and divergence time estimation. Mol Phylogenet Evol. 46:801–806.
Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10:1396–1401.
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, Yoder AD. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol. 52:705–716.
Zanis MJ, Soltis DE, Soltis PS, Mathews S, Donoghue MJ. 2002. The root of the angiosperms revisited. Proc Natl Acad Sci USA. 99:6848–6853.

Jeffrey Thorne, Associate Editor

Accepted August 9, 2009