Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution xxx (2007) xxx–xxx
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short Communication

An independent heterotachy model and its implications for phylogeny and divergence time estimation

Jihua Wu [a,*], Edward Susko [a], Andrew J. Roger [b]

[a] Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada B3H 3J5
[b] Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada B3H 4H7

Received 7 May 2007; revised 13 June 2007; accepted 29 June 2007

1. Introduction

Heterotachy or variation of rates-across-sites and lineages is widespread (e.g., Fitch and Markowitz, 1970; Uzzel and Corbin, 1971; Lopez et al., 2002) and can cause biased tree estimation (Lockhart et al., 1998; Susko et al., 2004; Inagaki et al., 2004). Kolaczkowski and Thornton (2004) described an intriguing heterotachy model that lead to topological bias for likelihood-based methods. That work generated a significant amount of discussion (e.g., Spencer et al., 2005; Lockhart and Steel, 2005; Gadagkar and Kumar, 2005) and raises questions about what forms of heterotachy might create these effects.

Here we consider a model that we will refer to as the random effects rate variation model (RERV[1]) which differs from common evolutionary models only in that the rate at a site is assigned independently to each edge in the tree according to a gamma distribution. In contrast, in the commonly used gamma rates-across-sites model of Yang (1993, 1994), each site has a different rate drawn from a gamma distribution but this rate is fixed for all edge lengths. The main heterotachy model that has been implemented in practice is the covarion or covariotide model described in Huelsenbeck (2002) and Galtier (2001). The model considered here similarly allows rates to vary across sites and lineages but has computational advantages. By comparison with covarion implementations, the model does not require a two-dimensional Markov chain implementation and compared with gamma rates-across-sites models, which use discretized versions of the gamma distribution, a full continuous gamma model can be implemented.

Fig. 1 illustrates the similarities and differences between trees estimated using the RERV model and a model with no rate variation (equal rates or ER) under the F81 or proportional model of Felsenstein (1981). The data set is a nucleotide alignment of 1530 sites from chloroplast-encoded psbB genes from 4 seed plants taken from Sanderson (2002). The likelihood for each of the three possible topologies is exactly the same under RERV and ER models; they thus both give the same estimate of topology. Still the edge lengths are dramatically different. We will show that the results in Fig. 1 are not coincidental. Under simple models of substitution like the proportional and Jukes and Cantor model (Swofford et al., 1996), the RERV model and ER models will give identical likelihoods. An implication of this is that no topological inconsistency is expected when a wrong ER model is used for data generated with heterotachy of the form described by the RERV model; maximum likelihood estimates of topologies will be identical. For more complex substitution models, however, likelihoods for RERV and ER need not agree. Still, we show that LogDet distances remain tree-additive, that is, there exist a set of edge lengths on the true topology, not necessarily interpretable as expected numbers of substitutions, that yield distances that match up with the LogDet distances. Significantly, in recent work describing phylogenetic artifacts due to heterotachy, the correlation of rates at sites for different lineages was very high. Our findings suggest that the presence of heterotachy with low levels of correlation between rates at different edges need not cause topological biases. However, the edge length differences illustrated in Fig. 1

* Corresponding author. Fax: +1 902 494 5130.
E-mail address: jihuaw@mathstat.dal.ca (J. Wu).

[1] *Abbreviations:* RERV, random effects rate variation model; ER, equal rates model; RAS, rate-across-sites model.

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2007.06.020

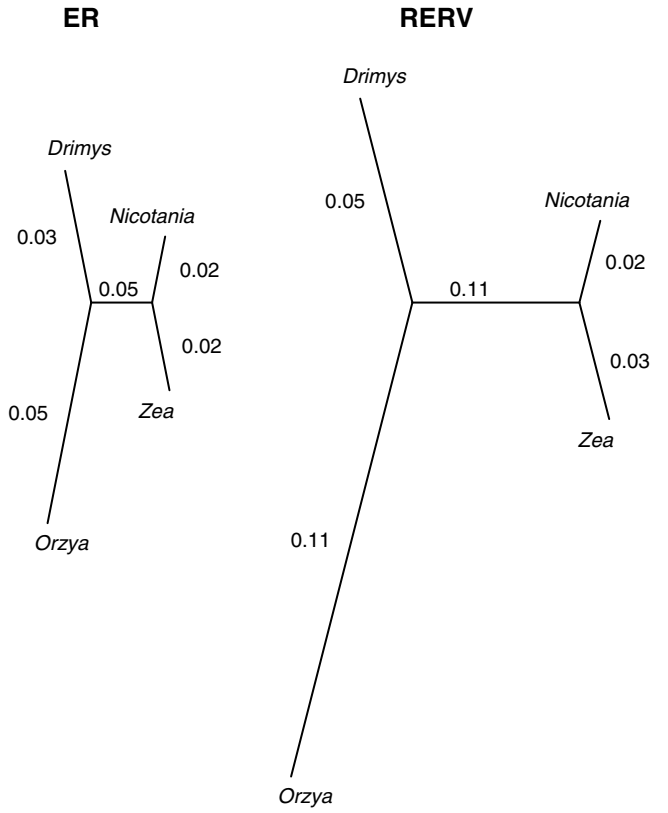

Fig. 1. The estimated trees for four of the taxa from the seed data under the ER and RERV ($\alpha = 0.11$) models. The log likelihoods were exactly the same for the two models.

are substantial. Differences like these will be shown to have a substantial effect on divergence time estimation.

2. Methods

The usual ER model assumes

1. Evolution in different sites is independent.
2. Conditional upon the character state (nucleotide or amino acid) at an internal node, character state evolution along adjacent edges is independent.
3. Evolution along an edge is according to a time-reversible Markov chain.

Let $p_{ij}(v)$ denote the probability that character state i is substituted by j along an edge of length v and let $P(v) = [p_{ij}(v)]$ be the substitution matrix. The substitution matrix is related to the rate matrix Q of the Markov model through matrix exponentiation $P(v) = \exp[Qv]$ and under time-reversibility, Q has a real-valued eigen-decomposition of the form $Q = SDS^{-1}$. In terms of this eigen-decomposition, the substitution matrix can be expressed as

$$P(v) = S\exp[Dv]S^{-1} \tag{1}$$

where the diagonal matrix $\exp[Dv]$ has iith entry $\exp[D_{ii}v]$. To ensure that the edge length v can be interpreted as the expected number of substitutions, Q is rescaled so that $-\sum_i \pi_i Q_{ii} = 1$.

The RERV model differs from the ER model in that for a given site h and edge e it independently assigns a rate multiplier r_{he} to that edge according to a gamma distribution with density $g(r;\alpha) = \alpha^{\alpha} r^{\alpha-1}\exp[-\alpha r]/\Gamma(\alpha)$. This allows evolution of character states to be faster or slower depending on the site and lineage. Given the rates, the model satisfies the properties 1–3 of the ER model with edge length v_e replaced by $r_{he}v_e$. Since rates are unobserved, the marginal substitution probabilities are obtained by integration with respect to the gamma distribution: $p_{ij}(v;\alpha) = \int p_{ij}(rv)g(r;\alpha)\,\mathrm{d}\alpha$. This integration can be done analytically and the substitution matrix $P(v;\alpha) = [p_{ij}(v;\alpha)]$ turns out to be

$$P(v;\alpha) = SD(v,\alpha)S^{-1}, \quad D(v,\alpha)_{ii} = (1 - D_{ii}v/\alpha)^{-\alpha} \tag{2}$$

Although, at each site, differing edge lengths are assigned to each edge, this assignment is at random from a distribution. It is only the parameter α in this distribution and the overall "average" edge lengths that require estimation. In other words, if we analyze a n species tree, we need to estimate $2n - 3$ branch lengths and a shape parameter α, $2n - 2$ in total.

The main other model of heterotachy implemented in practice is the covarion model (Tuffley and Steel, 1998; Huelsenbeck, 2002; Galtier, 2001). Rate multipliers r_{he} similarly vary over edges and sites, but switch stochastically over the tree making rates for a given edge dependent on the rates at parental edges. In contrast the RERV model independently assigns rates to edges and therefore has no memory of parental edge rates.

The covarion model is implemented through characterization as a two-dimensional Markov model. Rather than modeling evolution of character states, the evolution of character states and rates jointly throughout the tree are modeled. The pruning algorithm is applicable but with n_r possible rates, sums are over $n_c \times n_r$ states. Similarly as with the gamma rates-across-sites model, only a finite number n_r of possible states can be incorporated in this model; a continuous model cannot be implemented. In contrast, the RERV model has the attractive computational feature of allowing a continuous gamma implementation. The reason that the covarion and gamma rates-across-sites models can only deal with finitely many rates is that the pruning algorithm, necessary for feasible computation of likelihoods at a site, requires that probabilities of character states at both internal and external nodes be expressed as a product of the probability π_j of character state j at a "root" node and a sequence of transition probabilities $p_{ij}(v)$, one for each edge. Conditional upon the rate r at a site, a similar product expression can be obtained for the gamma model rates-across-sites model (with $p_{ij}(v)$ replaced by $p_{ij}(rv)$) but computing the unconditional probability requires integration over the common rate r and the result is a probability that can no longer be expressed as a product. In contrast to the gamma model, in the RERV model, separate rates are

assigned independently to edges so that integration is performed separately for each edge, thus preserving the product form required for the pruning algorithm.

To illustrate the improvements in likelihood possible through the RERV model, we consider a subset of the elongation factor 1α data considered previously in Susko et al. (2003). This amino acid data set consisted of 5 eukaryotes, 5 archaebacteria and 349 sites. Comparisons to the ER model give a likelihood ratio statistic

$$2\{\log\ \text{lik}[\text{RERV}] - \log\ \text{lik}[\text{ER}]\} = 17.74$$

This is much larger than the conventional cutoff, given by the 95th percentile of a χ^2_1 distribution, of 3.64 and corresponding to a 0.05 level of significance. In contrast to the original nucleotide data example, we see that for amino acid data there is some ability to distinguish between the RERV and ER models. In contrast to the original nucleotide data example, we see that for amino acid data there is some ability to distinguish between the RERV and ER models. With RAS variation added in to both models, the gap between the fits of the models becomes less. Letting $\text{JTT}+\Gamma$ denote the discrete gamma model with four categories of rates described in Yang (1994) and $\text{RERV}+\text{JTT}+\Gamma$ the RERV model with an additional overall rate multiplier at each site drawn from a discrete gamma model with four categories of rates, we obtained

$$2\{\log\ \text{lik}[\text{RERV}+\text{JTT}+\Gamma] - \log\ \ \text{lik}[\text{JTT}+\Gamma]\} = 4.42$$

which is only slightly larger than the conventional cutoff of 3.64 corresponding to a 0.05 level of significance.

The situation changes when we consider simple nucleotide substitution models. We will show that under the Jukes and Cantor and the proportional model, there are edge lengths such that the probabilities of terminal node data will be the same for the ER and RERV models. Since differences in the probabilities of data observed at terminal nodes are required in order to distinguish between two models, we can see that no matter how much data is accumulated, for these substitution models, there is no way of distinguishing between the ER and RERV models.

Consider two different α parameters, α and α′ for the RERV model; the ER model (1) is a special case of the RERV with $\alpha = +\infty$. Given a fixed topology, suppose that for each edge with edge length v an alternative edge length v' can be chosen so that

$$P(v;\alpha) = P(v';\alpha') \tag{3}$$

Since rates are assigned independently to edges, if (3) holds for each edge in the tree, the pruning algorithm will give the same probabilities for the data observed at terminal nodes. Because of the common S in (2), (3) reduces to $D(v,\alpha) = D(v',\alpha')$ or

$$(1 - D_{ii}v/\alpha)^{-\alpha} = (1 - D_{ii}v'/\alpha')^{-\alpha'}, \quad i = 1,\ldots,n_c \tag{4}$$

where n_c is the number of character states; 4 for nucleotide models and 20 for amino acid models. The Jukes and Cantor and proportional rate matrices have a single non-zero eigenvalue, $D_{ii} = -[\sum \pi_j(1-\pi_j)]^{-1} =: d$, giving the solution of (4) as

$$v' = \lim_{\alpha\to\infty}(\alpha'/d)[1 - \{(1 - dv/\alpha)^{\alpha}\}^{1/\alpha'}] = (\alpha'/d)[1 - \mathrm{e}^{-dv/\alpha'}] \tag{5}$$

Thus, given edge lengths v for an ER model, for the same topology but with edge lengths v' in (5), the likelihood under the RERV model will always coincide with the likelihood under the ER model. In contrast to the Jukes and Cantor and proportional models, most amino acid models have 19 distinct non-zero eigenvalues in their rate matrices. The $-D_{ii}$ is related to the rate at which character state i is substituted. We see here that the differences between these rates are crucial to inference about rate variation.

Even for amino acid data, where RERV and ER models will not give identical likelihoods, failure to adjust for real RERV heterotachous variation need not lead to inconsistent topological estimation. We illustrate this by showing that LogDet distances (Steel, 1994; Lockhart et al., 1994; Lake, 1994) converge upon tree-additive distances in the case where the data were generated by the RERV process. As sequence lengths get large, the LogDet distance, d_{rs}, between taxa r and s becomes approximately equal to $-\log\det[F^{(r,s)}]$, where $F^{(r,s)}$ is a matrix with ijth entry giving the probability of character states i and j for taxa r and s, respectively. These distances will be tree-additive if they can be expressed as the sum of some non-negative edge lengths v'_e, summed over all edges in the path between r and s; the $v'_e > 0$ need not be edge lengths actually present in the tree. To illustrate this consider the tree, in Newick format, ((1:v_1, 2:v_2):v_5, (3:v_3, 4:v_4)).

We restrict attention to d_{13} and show that it can be expressed as $v'_1 + v'_5 + v'_3$ for an appropriate choice of v'_i. The other distances can similarly be shown to be sums of the v'_i. Tree additivity on a four taxon tree suffices for consistent estimation since, with larger trees, knowledge of all embedded 4-taxon trees can be used to infer the larger tree.

The key here is that $F^{(1,3)}$ has the same product form as in the ER case:

$$F^{(1,3)} = \Pi P(v_1;\alpha)P(v_5;\alpha)P(v_3;\alpha) \tag{6}$$

Since the log of the determinant of a product of matrices is the sum of the log of the determinants of the individual matrices we obtain that

$$\begin{aligned} d_{13} &= -\log\det[F^{(1,3)}] \\ &= -\log\det[\Pi] - \log\det[P(v_1;\alpha)] - \log\det[P(v_5;\alpha)] \\ &\quad - \log\det[P(v_3;\alpha)] \end{aligned} \tag{7}$$

The contributions $-\log\det[P(v_1;\alpha)]$ can be shown to equal $\alpha\sum_j \log[1 - d_{jj}v_i/\alpha] > 0$, so that setting $v'_5 = -\log\det[P(v_5;\alpha)]$ and $v'_i = -\log\det[P(v_i;\alpha)] - 0.5\log\det[\Pi]$ for terminal edges we can see that (7) can be alternatively expressed as $d_{13} = v'_1 + v'_5 + v'_3$.

While we have established that consistent topological estimation can still be achieved while ignoring the presence

of RERV, the example in Fig. 1 illustrates that the inability to distinguish between data generated under an RERV model and data generated under a ER model can lead to substantial errors in edge length estimates if the wrong model is assumed. These will, in turn, affect divergence time estimation. We illustrate this using one of the common methods for divergence time estimation: non-parametric rate smoothing (NPRS), described in Sanderson (1997) and implemented in the program r8s described in Sanderson (2003). The program takes as input edge lengths estimated edge lengths v_k, and chooses divergence times T_k so that neighboring rates, estimated through $r_k = v_k/T_k$ are similar to each other. We ran the r8s program with default settings on a 13 sequence subset of the psbB seed plant data from Sanderson (2002). For maximum likelihood analysis, we used a proportional substitution model (F81) and estimated edge lengths from the ER model as well as the RERV model with $\alpha = 0.028$. As indicated earlier, the RERV model and the ER model give exactly the same likelihood so there is no way of distinguishing which gives better edge lengths. The estimated tree is given in Fig. 2. A summary of the estimated divergence times under ER and percent changes when edge lengths are estimated under RERV is given in Table 1. The most obvious differences between the time estimates are that, relative to the ER model, the shallowest nodes are estimated to be much younger by the RERV model and the deepest nodes are much older under RERV. The differences observed are substantial, with a maximum age estimate difference of >90 million years for node 13.

Table 1
The estimated divergence times for the seed plant data

Node	Estimated divergence time		Percent change (%)
	ER	RERV	
[*] LP	450.00	450.00	0.0
(2)	389.71	423.52	8.0
(3)	218.70	276.23	20.8
(7)	395.64	435.13	9.1
(8)	371.38	429.62	13.6
(9)	173.72	195.77	11.3
(13)	315.33	406.18	22.4
(16)	216.28	129.39	−67.2
(17)	181.16	119.98	−51.0
(19)	119.34	95.27	−25.3
(22)	56.31	15.10	−272.9

Edge lengths under the ER and RERV model ($\alpha = 0.028$) were input to the r8s program with default settings. The node labels correspond to the nodes in Fig. 2.

3. Discussion

Due to the ease of computational implementation, the RERV model considered here has the potential to be useful in making inferences about heterotachy. Under it, posterior estimates of rates at a site in different portions of the tree can be allowed to vary which could prove useful in analyzing how functional constraints might be changing across sites and lineages.

Under the RERV model, topological inconsistency was avoidable through the use of LogDet distances. Distance methods combined with LogDet distance have the

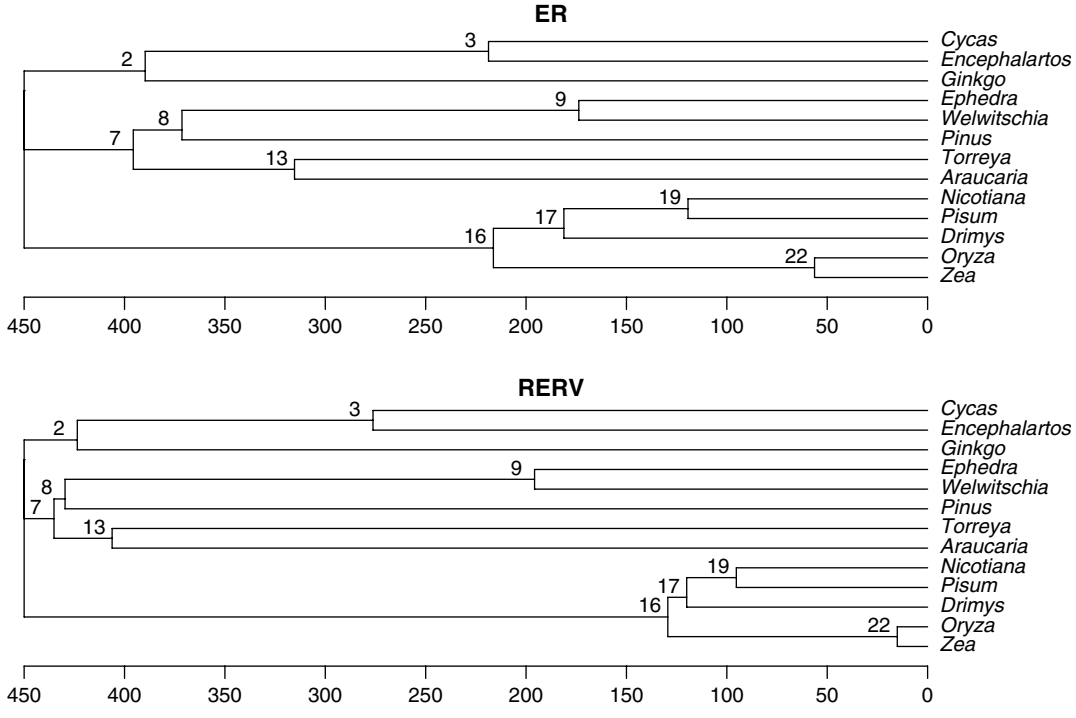

Fig. 2. The estimated trees for the seed data with divergence times from the r8s program under both the ER and RERV models.

advantage that they do not require the knowledge that we were dealing with a heterotachous model. Like all distance methods, however they use the data only through pairwise relationships, and with amino acid data suffer from sparseness difficulties (often LogDet distances for a pair do not exist because one or more character state pairs were not observed).

The LogDet consistency result here is inline with Susko et al. (2004) who found that inconsistent topological estimation did result when heterotachous parallel rate variation was occurring but vanished as the correlation in rates at edges became negligible. Under the RERV model here, rates are independently assigned to edges; the correlation of rates is zero. Similarly in Kolaczkowski and Thornton (2004), inconsistent topological estimation using a standard homotachous model resulted if the data were generated by a heterotachous model with perfectly correlated rates on edges. The argument above that LogDet distances remain tree-additive with an RERV generating model relied mainly on the fact that the heterotachy considered here was independent and generalizes more broadly to differing distributions of rates that maintain this property. However, more investigation of heterotachy effects is required before definitive statements can be made about how much correlation can be tolerated and whether results extend to other estimation settings.

Much of the work on effects of model misspecification has focused on topological estimation. Divergence time estimation is an enterprise of substantial interest as well. Variation of average rates, or molecular clock violations, have long been recognized as an issue that requires correction. Here we see that heterotachy, a different form of rate variation, can similarly have substantial effects on divergence time estimation. With more complex substitution models, there is potential to distinguish between heterotachous and non-heterotachous models, but as the non-identifiability under the proportional model illustrates, in some cases the presence of heterotachy can render the problem of divergence time estimation insoluble.

Due to the tertiary structure of protein, codependency of site rates is very common. More realistic models make adjustments in order to deal with site codependency. For instance, Stern and Pupko (2006) gave an evolutionary space–time model with varying among-site dependencies. Similar work was done by Fares and McNally (2006). Most such models require structural information, however. Although the model considered here is as an independence model, it might alternatively be thought of as a probabilistic model for the marginal distributions of site patterns given only the data at the site. One will obtain consistent estimation for dependent data when using an independence model likelihood if the marginal distribution of data at a site are correctly modeled. Adjustment can then usually be made in a similar fashion to the original argument of Wald (1949), replacing uses of the law of large numbers for independent observations with dependence analogues.

Acknowledgments

We thank two anonymous reviewers for helpful comments. This research was supported by Discovery grants awarded to E.S. and A.J.R. by the Natural Sciences and Engineering Research Council of Canada. A.J.R. and E.S. are fellows of the Canadian Institute for Advanced Research Program in Evolutionary Biology. A.J.R. is supported by an NSERC E.W.R. Steacie fellowship and the Peter Lougheed New Investigator Award from the Canadian Institutes of Health Research and the Peter Lougheed Medical Research Foundation.

References

Fares, Mario A., McNally, David, 2006. CAPS: coevolution analysis using protein sequences. Bioinformatics 22, 2821–2822.

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

Fitch, W.M., 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.

Gadagkar, S.R., Kumar, S., 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol. Biol. Evol. 22, 2139–2141.

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

Huelsenbeck, J.P., 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19, 698–707.

Inagaki, Y., Susko, E., Fast, N.M., Roger, A.J., 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF1-α phylogenies. Mol. Biol. Evol. 21, 1340–1349.

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

Lake, J.A., 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA 91, 1455–1459.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Charleston, M.A., Howe, C.J., 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15, 1183–1188.

Lockhart, P.J., Steel, M.A., Hendy, M.D., Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605–612.

Lockhart, P.J., Steel, M.A., 2005. A tale of two processes. Syst. Biol. 54, 948–951.

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

Sanderson, M.J., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14, 1218–1231.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.

Sanderson, M.J., 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19, 301–302.

Spencer, M., Susko, E., Roger, A.J., 2005. Likelihood, parsimony and heterogeneous evolution. Mol. Biol. Evol. 22, 1161–1164.

Steel, M.A., 1994. Recovering a tree from the Markov leaf colourations it generates under a Markov model. Appl. Math. Lett. 7, 19–23.

Stern, A., Pupko, T., 2006. An evolutionary space-time model with varying among-site dependencies. Mol. Biol. Evol. 23 (2), 392–400.

Susko, E., Field, C., Blouin, C., Roger, A.J., 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. Syst. Biol. 52, 594–603.

Susko, E., Inagaki, Y., Roger, A.J., 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol. Biol. Evol. 21, 1629–1642.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D., Moritz, C., Mable, B. (Eds.), Molecular Systematics, second ed. Sinauer Associates, Sunderland, MA, pp. 407–514.

Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147, 63–91.

Uzzel, T., Corbin, K.W., 1971. Fitting discrete probability distributions to evolutionary events. Science 173, 1089–1096.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimator. Ann. Math. Stat. 20, 595–601.

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.

Please cite this article in press as: Wu, J. et al., An independent heterotachy model and its implications for ..., Mol. Phylogenet. Evol. (2007), doi:10.1016/j.ympev.2007.06.020