A Test for Heterotachy Using Multiple Pairs of Sequences

Jihua Wu*,1 and Edward Susko1

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

*Corresponding author: E-mail: jwu@ms.soph.uab.edu.

Associate editor: Hervé Philippe

Abstract

Heterotachy is a general term to describe positions that evolve at different rates in different lineages. Heterotachy also 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 (Wu J, Susko E. 2009. General heterotachy and distance method adjustments. Mol Biol Evol. 26:2689–2697). Motivated by this result, we propose three new distance-based tests: a heterogeneity test, a heterotachy test, and a within-gene heterotachy test and demonstrate with simulations that they perform well under a wide range of conditions. We also applied the first two tests to two real data sets and found that although all these data sets showed significant evidence of heterotachy, there were subtrees for which the data were consistent with an equal rates or rates-across-sites model.

Key words: heterogeneity, heterotachy, within-gene heterotachy, covarion model, distance method, hypothesis test.

Introduction

The term heterotachy (meaning “different speeds”) was introduced by Lopez et al. (2002) to refer to the situation in which rates vary across both sites and lineages and is common in protein evolution (Galtier 2001; Huelsenbeck 2002; Wang et al. 2007). Several special cases of heterotachy models were proposed. Covarion models were originally described as biological models in Fitch and Markowitz (1970) but the term has more recently been used to describe a class of heterotachy models where rates at a site switch according to a stationary Markov chain model throughout the tree (Tuffley and Steel 1998; Galtier 2001; Huelsenbeck 2002; Wang et al. 2007). Susko et al. (2003) propose a bivariate rates model in which rates in two subtrees may be different. Kolaczkowski and Thornton (2004) (K&T) also described an interesting heterotachy model where parsimony outperforms the misspecified conventional likelihood-based method that makes no adjustment for heterotachy; that parsimony does not generally outperform uncorrected likelihood methods in dealing with heterotachy has since been shown by a number of authors (cf. Gadagkar and Kumar 2005, Philippe, Zhou, et al. 2005; Spencer et al. 2005). In contrast, for the random effect rates variation (RERV) model (Wu et al. 2007) where rates vary 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. Wu and Susko (2009) gave a general definition of heterotachy as multivariate rates-across-sites variation that includes the equal rates (ERs) model, rates-across-sites (RASs) model, covarion model, bivariate model, and RERV model as its special cases. They also proposed a pairwise alpha heterotachy adjustment (PAHA) method and showed that it can give much improved topological estimation in the K&T four-taxon model (Kolaczkowski and Thornton 2004, 2008). The present article supplements this previous work with tests for heterotachy based on estimates coming from the PAHA method.

Recently, a few formal tests of covarion evolution have been published (e.g., Lopez et al. 1999; Misof et al. 2002; Susko et al. 2002; Baеle et al. 2006). Lockhart et al. (1998) proposed two nonparametric tests: a contingency test and an inequality test. The null hypothesis of the inequality test is that the sequences follow a RAS model, and the alternative hypothesis is that the sequences follow covarion evolution as defined by Shoemaker and Fitch (1989). Galtier (2001) used approximate likelihood estimates to perform the likelihood ratio test (LRT), and Huelsenbeck (2002) proposed a integrated LRT for covarion evolution using Markov chain Monte Carlo. Though LRTs often perform well in evolutionary model testing (Posada and Crandall 2000), the computations are expensive and difficult to implement. Ané et al. (2005) also proposed a nonparametric test called the covarion test in their paper. Its null hypothesis is the same as the inequality test proposed by Lockhart et al. (1998), but the alternative hypothesis is that the sequences follow a model of both among-site rate variation and covarion evolution. However, these tests (Lockhart et al. 1998; Ané et al. 2005) depend on the tree topology, and the tree needs to be split into several groups. Moreover, these tests use the information in the sequences only through counts of site patterns where some subsets of taxa are invariable. Zhou et al. (2010) proposed posterior predictive discrepancy tests, two of which (D H test and D 0 test) can detect heterotachy signal in the data set. However, these two tests are designed to assess their covarion mixture model, and the null distributions of the tests are dependent on parameters of unknown value. So they might not be convenient in heterotachy test. Wang et al. (2007) proposed a general heterotachy model that combines features of both models proposed by Galtier (2001) and
This nesting of models allows for LRTs. They use the software PROCOV (Wang et al. 2007) that we refer to as the PROCOV test. The test is actually a set of LRTs for testing the null hypotheses of a RAS model against the alternatives of a Galtier, Huelsenbeck, or general heterotachy model. For the PROCOV test, the extensive computation of the likelihood is unavoidable, so for large data sets, a new method need to be developed. Such tests are also intended to detect evidence for a covarion model, which is only a particular type of heterotachy.

Wu and Susko (2009) showed that heterotachy can generally be viewed as RAS variation for pairs. Motivated by this idea, they proposed PAHA that allows rates between two taxa to follow different distributions (a gamma distribution with different shape parameter is used for each pair of taxa); maximum likelihood estimation is used to find the best distance and the best (shape) parameter $\alpha$ for the corresponding distribution for each pair of taxa. In this article, we extract information from $\alpha$s calculated through PAHA method (Wu and Susko 2009) and construct a test statistic to decide whether the RAS model or general heterotachy model is appropriate for a group of sequences. If it is true that the underlying $\alpha$ parameters are the same for each pair, a RAS model is appropriate. The heterotachy test we constructed is effectively a distance between the $\alpha$ estimates that adjusts for their covariance and has a known chi-squared distribution under the null hypothesis of a RAS model. Similar ideas can be used to construct a heterogeneity test that can be used to distinguish ER model from RAS model. We examine the performance and power of the two new tests under a wide range of conditions using simulations. We also use these two tests to analyze three real data sets.

Concatenated data sets with large numbers of sites are commonly constructed by concatenating gene alignments. For such alignments, differing edge-lengths for differing genes is a form of heterotachy that we refer to as between-gene heterotachy. Because there are many genes in these data sets, this raises a question: where does the heterotachy come from? Does heterotachy occur within the genes or between the genes or both? A number of recent studies indicate that failing to adjust for between-gene heterotachy can lead to erroneous conclusions. For instance, in a study on eutherian mammals, Nishihara et al. (2007) found Afrotheria and Xenartha were erroneously grouped together in a nucleotide analysis of the concatenated data, whereas analyses that allowed different edge-lengths for different genes supported a correct placement. Similarly, in considering plastid genes, Rodriguez-Ezpeleta et al. (2007) found that in a concatenated analysis, Mesostigma was not located at the base of the Streptophyta tree in contrast to analyses with nuclear and mitochondrial genes. Partitioning the plastid data into three functional groups gave one group with strong support for Mesostigma at the base of the Streptophyta and two others with weak support for alternative placements. Som and Fuellen (2009) showed that the neighbor joining method (Saitou and Nei 1987) is affected when evolutionary rates in a concatenation exhibit heterotachy. They proposed a simple approach for inferring more accurate multigene species phylogenies under heterotachous conditions. Rasmussen and Kells (2007) found that gene trees exhibit common properties that can be exploited for evolutionary studies and accurate phylogenetic reconstruction. They decoupled evolutionary rates into gene-specific and species-specific components.

Using the PAHA method proposed in Wu and Susko (2009), we can adjust between-gene heterotachy by assuming different gene trees for different genes. Any additional heterotachy must be within-gene heterotachy, which should result in differing shape parameter $\alpha$s. We propose a within-gene heterogeneity test $T_2$ and show that $T_2$ is approximately chi-square distributed under the null hypothesis. The results of simulations show that $T_2$ performs well under a wide range of conditions. We also applied the test to chloroplast genome sequences studied by Gruenheit et al. (2008).

**Methods**

Wu and Susko (2009) proposed two equivalent definitions for the general heterotachy and showed that a variety of heterotachy models are all special cases of the general heterotachy model. These include the ER model, RAS model (Yang 1993), K&T four-taxon model (Kolaczkowski and Thornton 2004), bivariate rates model (Susko et al. 2003), RERV model (Wu et al. 2007), and covarion model (Tuffley and Steel 1998). Wu and Susko (2009) also showed that, for pairs of taxa, heterotachy is simply RAS variation with different RAS distributions for different pairs. For heterotachy, the implied (marginal) distribution for pairs will vary from pair to pair, but for RAS, it will be the same for each pair. In this generalized heterotachy model, for each pair of taxa, heterotachy can be modeled as RAS variation for pairs. Motivated by this result, we assume that rates between two taxa follow a gamma distribution $R(r, \alpha)$ (Yang 1993, 1994; Tuffley and Steel 1998) with a different shape parameter $\alpha$ for different pair of taxa.

We estimate separate $\alpha$s for pairs through maximum likelihood (ML) estimation, then find the best (shape) parameter $\alpha$ for each pair of taxa. Generally, all tests involve checking whether $\alpha$s satisfy conditions implied by the null hypothesis after adjusting for uncertainty in estimation. If the true underlying $\alpha$s are all infinite, the sequences follow an ERs model. In the general heterotachy model that corresponds to RAS, the RAS distributions for each pair is the same and so it is to be expected that similar alphas will be estimated for each pair. In more general heterotachy models that are not also RAS models, it will usually be the case that pairwise distributions differ. Our alternative hypothesis is thus always that the alphas are different for at least some pairs of taxa. It should be pointed out, however, that there are special exceptional heterotachy models where this is not the case. An example is the usual covarion model (Tuffley and Steel 1998) on a star tree with all equal edges. Because of the stationarity of the Markov models used for rates, the pairwise distributions will be the same. It is convenient to
define $\hat{\beta}_i = \alpha_i^{-1}$ for the $i$th pair of taxa so that the ER model corresponds to $\hat{\beta}_i = 0$. We let $\tilde{\beta}_i = 1/\hat{\beta}_i$ denote the ML estimate of $\beta$, and construct our test statistics using $\tilde{\beta}$, where $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_n)$ and $n_p$ is the number of pairs. For instance, with four taxa $n_p = 6$. All the test statistics have the form of $b^T\Sigma^{-1}b$, where vector $b$ is a function of the pairwise $\beta$s, and the matrix $\Sigma$ is chosen to be an estimation of covariance matrix so that resulting test statistics are $\chi^2$ distributed.

For a group of taxa, suppose that there are $g$ ($g \geq 2$) genes for each taxa. From Definition 2 in Wu and Susko (2009), heterotachy can be considered as multivariate edge-length variation across sites and lineages. Thus, with $g$ genes, heterotachy can arise because these genes have different edge-lengths, a case that we refer to as between-gene heterotachy. This could be true whether or not there is additional variation across sites and lineages within genes. Whether such additional within-gene heterotachy exists is of interest. Because boundaries between genes are known, between-gene heterotachy can be adjusted for by allowing different edge-lengths for different genes. We can adjust for between-gene heterotachy by assuming different gene trees for different genes. Any additional heterotachy detected after adjustment must be within-gene heterotachy. The within-gene heterotachy test statistic is similar to the heterotachy test statistic. As with the test for heterotachy, under the null hypothesis of RAS, the $\alpha$ for a pair will be the same for different genes. However, the distances $d$ are allowed to be different for different genes.

For a pair of taxa, denote the log likelihood for gene $u$ as $l(\alpha, d_u)$. Then the log likelihood for all the genes will be the sum (i.e., $\sum_{u=1}^{g} l(\alpha, d_u)$). Through maximum likelihood estimation, we find the best (shape) parameter $\alpha$ for each pair of taxa. Because the between-gene heterotachy was adjusted by assuming different gene trees with different branch lengths, any additional heterotachy must be a consequence of within-gene heterotachy, the information of which was contained in $\alpha$s. If the true underlying $\alpha$s for all the pairs are the same, there is no within-gene heterotachy for all the genes; if the $\alpha$s are different, the group of taxa may be experiencing within-gene heterotachy. We will construct an approximately chi-squared test statistic to check whether the $\alpha$s satisfy one of these conditions.

Three Test Statistics

The tests considered differ in their null hypothesis but have an alternative hypothesis that includes any violation of the null hypothesis. In the heterogeneity test, the null hypothesis is that the sequences follow the ER model in which the rates for all pairs and sites are equal. This means that the $\alpha$s are all equal to infinity, which is equivalent to the $\beta$s being equal to 0:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{n_p} = 0.$$ 

Let $b_1 = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_n)$. Then the test statistic is

$$T_1 = b_1 \Sigma_{1}^{-1}b_1^T,$$  \hspace{1cm} (1)

where $\Sigma_{1}$ is the covariance matrix for vector $b_1$ and will be derived in Appendix A1 (see Supplementary Material online). $T_1$ should be approximately chi-square distributed with $n_p$ degrees of freedom (see Appendix A1 in Supplementary Material online for details).

In the heterotachy test, the null hypothesis is that the sequences follow a RAS model in which the rates for all pairs are drawn from the same gamma distribution. This means that all the $\alpha$s or equivalently all the $\beta$s are equal:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{n_p}.$$ 

We denote the mean of all the $\hat{\beta}$s as $\bar{\hat{\beta}} = \frac{\sum_{i=1}^{n_p} \hat{\beta}_i}{n_p}$, and $b_2 = (\beta_1 - \bar{\hat{\beta}}, \beta_2 - \bar{\hat{\beta}}, \ldots, \beta_{n_p} - \bar{\hat{\beta}})$. Then the test statistic is

$$T_2 = b_2 \Sigma_2^{-1}b_2^T,$$ \hspace{1cm} (2)

where $\Sigma_2$ is the covariance matrix for vector $b_2$. The covariance matrix of $(\hat{\beta}_1 - \bar{\hat{\beta}}, \hat{\beta}_2 - \bar{\hat{\beta}}, \ldots, \hat{\beta}_{n_p} - \bar{\hat{\beta}})$ is not invertible. For this reason, we did not use $\beta_{n_p} - \bar{\hat{\beta}}$ when we construct the test statistic $T_2$. Thus, $T_2$ should be approximately chi-square distributed with $n_p - 1$ degrees of freedom (see Appendix A1 in Supplementary Material online for details). The heterotachy test requires estimation of the covariance matrix $\Sigma_2$, which is derived in Appendix A1 (see Supplementary Material online).

In the within-gene heterotachy test, the null hypothesis is that the sequences follow a RAS model and possibly have between-gene heterotachy but no within-gene heterotachy. Between-gene heterotachy can be adjusted for by assuming different gene trees for different genes, so even if between-gene heterotachy exists, as long as there is no within-gene heterotachy, the estimated $\hat{\beta}$s (or $\alpha$s) should not show a significant difference. The null hypothesis,

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{n_p},$$

is the same as for $T_2$. The within-gene heterotachy test statistic is

$$T_3 = b_3 \Sigma_3^{-1}b_3^T,$$ \hspace{1cm} (3)

where $b_3$ is the same as $b_2$. $T_3$ looks similar to $T_2$, but $\Sigma_2$ and $\Sigma_3$ are different, though they are calculated using similar ideas. In Appendix A2 (see Supplementary Material online), we will derive an approximate covariance matrix $\Sigma_3$ and prove that $T_3$ is approximately chi-square distributed with $n_p - 1$ degrees of freedom.

Simulation Settings

The performance of the three tests were examined with simulation experiments. To examine the type 1 error rate (the level) of the three tests or how often the test rejects the null hypothesis when it is true, 1,000 replicate alignments under null hypothesis were simulated using the Jones, Taylor, and Thornton (JTT) model through seq-gen (Rambaut and Grassey 1997).

- For the heterogeneity test $T_1$: under ER model and tree in Newick format ($(A : 0.75, B : 0.05) : 0.1, (C : 0.75, D : 0.05)$).
• For the heterotachy test $T_2$: under gamma-distributed rate heterotachy (RAS model) with a shape parameter $\alpha$ ($\alpha = 0.25$ unless otherwise noted) and tree ($(A : 0.75, B : 0.05) : 0.1, (C : 0.75, D : 0.05)$).

• For the within-gene heterotachy test $T_1$: under gamma-distributed rate heterotachy (RAS model) with a shape parameter $\alpha$ ($\alpha = 0.25$ unless otherwise noted) for each of the two gene trees ($(A : 0.75, B : 0.05) : 0.1, (C : 0.75, D : 0.05)$) and ($(A : 0.05, B : 0.75) : 0.1, (C : 0.05, D : 0.75)$). The two genes were of equal length unless otherwise noted, then concatenated sequences generated from each gene into one alignment.

To examine the power of the heterogeneity test or how often it correctly rejects the null hypothesis, 1,000 replicate alignments were simulated under the alternative hypothesis as follows:

• For the heterogeneity test $T_1$: under gamma-distributed rate heterotachy (RAS model) with a shape parameter $\alpha$ ($\alpha = 0.25$ unless otherwise noted) and the same tree as in null hypothesis through seq-gen.

• For the heterotachy test $T_2$: 1) under a K&T four-taxa model. Sites were randomly assigned to one of two trees: ($(A : 0.75, B : 0.05) : 0.1, (C : 0.75, D : 0.05)$) and ($(A : 0.05, B : 0.75) : 0.1, (C : 0.05, D : 0.75)$). Note that this differs from the null within-gene model in that, for instance, sites generated from the first tree need not all be from the same gene. Additionally, rates for sites were generated from a gamma RAS model with $\alpha = 0.25$ unless otherwise specified; 2) under a covarion model of Tuffley and Steel (1998), with “ON” frequency 0.25 and speed of covarion evolution 1.75, along a tree, ($(A : 0.75, B : 0.05) : 0.1, (C : 0.75, D : 0.05)$) through seq-gen-aminocov (Wang et al. 2007), which is available at http://www.liv.ac.uk/~matts/covarion.html.

• For the within-gene heterotachy test $T_3$: under Tuffley and Steel (1998) with ON frequency 0.25 and speed of covarion evolution 1.75 for each of the two gene trees ($(A : 0.75, B : 0.05) : 0.1, (C : 0.75, D : 0.05)$) and ($(A : 0.05, B : 0.75) : 0.1, (C : 0.05, D : 0.75)$) through seq-gen-aminocov. The two genes were of equal size unless otherwise noted. We simulated two different kinds of sequences, one was COV + Gamma (with $\alpha = 0.25$ unless otherwise specified), and the other was just COV without gamma.

We varied several conditions to evaluate the sensitivity of the tests. First, the effect of sequence length on the level and the power of the tests was examined. Additionally, we examined the effect of the RAS gamma distribution parameter (shape parameter $\alpha$) on the level and the power of the tests. To examine the effect of the shape parameter $\alpha$ on $T_1$, simulations were conducted with $\alpha$ values ranging between 0.25 and 20 and with sequence lengths of 2,000 and 4,000. To examine the effect of the shape parameter $\alpha$ on $T_2$, simulations were performed with $\alpha$ values of 0.25, 0.5, 0.75, 1.0, and 1.25 and with a sequence length of 4,000. To examine the effect of the shape parameter $\alpha$ on $T_3$, simulations were performed with $\alpha$ values ranging between 0.05 and 1.5 and with a sequence length of 4,000. We also examined the effect of invariable sites for $T_1$ and $T_2$. Sequences with 4,000 sites were simulated with $\alpha = 0.25$ and 0%, 10%, 15%, 20%, and 25% of sites being invariant.

Finally, we changed the degree of heterotachy in the following two ways and examined its effect on the level and the power of $T_2$ and $T_3$.

1) We changed the longest branch length in the two partitions from 0.75 to 0.05 by 0.1 in each case. Note that if the longest branch is changed from 0.75 to 0.05, then there is no heterotachy. Smaller choices of longest branch lengths correspond to lower levels of heterotachy.

2) We changed the proportion of sites coming from the first partition. The model moves away from a RAS model as the proportion of sites in the first partition goes from 0.0 to 0.5.

Results

Performance of Heterogeneity Test in Simulation

From figure 1a, we can see that the probability of rejection under the null for the tests is close to or below the targeted level of 5%. When $\alpha = 0.25$ and sequence length is larger than 1,000, the power of the tests is 1.0. When the value of $\alpha$ is increased to 4.0, the percentage of rejections decreases. (Note that when $\alpha$ is large, generating from the RAS model is approximately the same as generating from an ER model). As expected, the power of the test increases with increasing sequence length. In figure 1b, under the null hypothesis, $\alpha$ is fixed at $+\infty$, so the level of the test is the same for all the simulations when $\alpha$ varies. In order to see how the power of the heterogeneity test changes, we plot the percentage of rejection according the value of 1/$\alpha$. When 1/$\alpha$ is very small (i.e., $\alpha$ is very large), the power of the test is as low as 0. This is reasonable because if $\alpha$ is very large, the RAS model will degrade to the ER model, which makes it very hard to distinguish between the null and alternative hypothesis.

The heterogeneity test seems as if it cannot differentiate the ER model from the RAS model when there are invariable sites both under the null hypothesis and the alternative hypothesis: If there are no invariable sites, the heterogeneity test performs perfectly with level 0.0 and power 1.0, but if even just 5% of the sites are invariable, the level of the test is close to 0.4 and the power of the test is 1.0. This is not completely a surprise because the ER model with some percentage of invariable sites is a kind of RAS model.

Performance of the Heterotachy Test in Simulation

The first set of simulations (fig. 2a–c) were under the K&T four-taxa model as described in the Methods section. The heterotachy test appears to perform well under a wide range of simulation conditions. The level of the test is close to or below the targeted level of 5% in each group. The power of the test increases with increasing sequence length (fig. 2a). In particular, when sequence length is larger than 5,000, the power is close to 1. The test’s power increases with increasing shape parameter $\alpha$ (fig. 2b). The power decreases with increasing proportions of invariable sites (fig. 2c).
In order to show that the heterotachy test can also distinguish between covarion model and a RAS model, we simulated amino acid sequences under the alternative hypothesis along a four-taxon tree in Newick format, 

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

The heterotachy test appears to perform well under a wide range of simulation conditions. The level of the tests is close to or below the targeted level of 5% in each group (fig. 2d–f). The power increases with increasing sequence length (fig. 2d). In particular, when sequence length is larger than 5,000, the power is close to 1. The power increases with increasing shape parameter $\alpha$ (fig. 2e). The power decreases with increasing proportions of invariable sites (fig. 2f).

In figure 3a and b, sequences under $H_0$ were simulated according to the K&T four-taxa model and sequences under $H_a$ were simulated according to the covarion model in figure 3c. If the longest branch length in the two trees is 0.05, the two trees are the same and there is no heterotachy. As this longest edge increases, the degree of heterotachy increases as well. From figure 3a, we can see that if the longest branch length in the two partitions changes from 0.05 to 0.75, the power increases as well. Similarly in figure 3b, if the proportion of sites in the first partition increases from 0.1 to 0.5, the degree of heterotachy increases and we find that the power will increase except when the proportion is 0.5. In figure 3c, the performance of test $T_2$ under the covarion model is similar to that of in figure 3a, where sequences under $H_a$ were simulated according to K&T four-taxa model. The power of the tests is low for smaller largest edge lengths. The difficulty is that when the largest edge lengths get close to the smaller ones, the tree gets closer to a star tree with equal edge lengths. For such a star tree, because of the stationarity of the covarion model, the pairwise distributions of rates are the same for each pair and it is expected that estimated alphas across pairs will be similar.

**Performance of the Within-Gene Heterotachy Test in Simulation**

The within-gene test (fig. 4a–d) appears to perform well under a wide range of simulation conditions. The level of the tests is close to or below the targeted level of 5% in each group.

The power of the test increases with increasing sequence length (fig. 4a). For the data simulated under the COV model, when sequence length is larger than 4,000, the power is close to 1. However, for the data simulated under the COV + Gamma model (the line in the middle, see fig. 4a, c, and d), the power is not as large as that of the data...
simulated under the COV model (the line in the top, see fig. 4a, c, and d). The reduced power suggests that the COV + Gamma model is in some sense closer to the null hypothesis. In figure 4b, only sequences that follow the COV + Gamma model were simulated because we wanted to observe the performance of $T_2$ when $\alpha$ varies. The test’s power increases with increasing shape parameter $\alpha$ (fig. 4b). However, when $\alpha$ is close to 0, both the level and the power of the test is close to 0.

From figure 4c, we can see that if the longest branch length in the two partitions changes from 0.05 to 0.75, which implies that as the degree of between-gene heterotachy increases, the power increases as well. As with fig. 3c, the power gets small when the longest edge is short. This is because the tree gets closer to a star tree with equal edges when the largest edges get small, and such a star tree, with a covarion model, yields the same pairwise distributions of rates. In this case, the covarion model will degrade to the RAS model, so the test does not perform well in this circumstance (a similar phenomenon can be found in fig. 3c).

In figure 4d, for the data simulated under the COV model, if the proportion of sites in the first partition increases from 0.1 to 0.5, implying that the degree of between-gene heterotachy increases, we find that the power equals 1 except when the proportion is 0.5. For the data simulated under the COV + Gamma model, the power of test $T_3$ is very high (about 90%) when the proportion of sites in partition 1 is 0.1. It decreases when the proportion of sites in partition 1 increases from 0.1 to 0.5. Contrast this with figure 4b, which is just the opposite. This should not be surprising because $T_3$ is designed to adjust for between-gene heterotachy and to test within-gene heterotachy only.
FIG. 3. Results of the simulation study showing the effect of the longest branch length, the proportion of sites in the first partition on the level, and the power of the heterotachy test $T_3$. (a) The shape parameter $\alpha$ was fixed as 0.25 and sequence length was fixed as 4,000 (sequences under $H_1$ were simulated under $K&T$ four-taxa model); (b) the shape parameter $\alpha$ was 0.25 and sequence length was 4,000 (sequences under $H_1$ were simulated under $K&T$ four-taxa model); (c) the shape parameter $\alpha$ was fixed as 0.25 and sequence length was fixed as 4,000 (sequences under $H_1$ were simulated under covarion model).

When the proportion of sites in partition 1 increases from 0.1 to 0.5, within-gene heterotachy decreases and between-gene heterotachy increases, so the power of the test $T_3$ will decrease.

Real Data Analysis for the Heterogeneity Test and the Heterotachy Test

For all the real data sets that we considered, sites with missing data were deleted, and we were always able to reject the ER and RAS models. We expect that this observation generalizes and that for many data sets with an appreciable number of taxa, heterotachy can be accepted as the correct model (Lopez et al. 1999). In such cases, it is still of interest to identify the subtrees where ER and RAS models cannot be rejected.

ER/RAS/Heterotachy Subtree Searching

In the following, we describe our procedure for finding relatively small heterotachy subtrees.

Step 1: Start from a three taxa subtree. For example, in figure 5, we can start from G1.

Step 2: Apply $T_2$ to the current subtree.

Step 3: If in Step 2, the null hypothesis is rejected, which means this subtree is the smallest heterotachy subtree, then try another subtree and go back to Step 1. If all the subtrees are checked, then stop.

Step 4: If in Step 2, the null hypothesis is not rejected, add the closest taxa to this subtree. If the next branch to be added to the current subtree is a taxa, we add it and go back to Step 2. For example, in figure 5, if G1 is not rejected, add taxa *Saccharomyces cerevisiae* to G1 and go back to Step 2. If the next branch is a two-taxa subtree, we add its taxa to the current subtree and go back to Step 2 (e.g., considering fig. 5, if G2 is not rejected, *Drosophila melanogaster* and *Anopheles gambiae* are added to G2 and we go back to Step 2); if the next branch is a subtree with more than two taxa, we need to apply the recursive procedure to this subtree. Assuming it is found to be an RAS subtree, we add its taxa to the current subtree and go back to Step 2.

The procedure for finding relatively large RAS/ER subtrees is similar to the above. Before applying the procedure to find the largest RAS/ER subtree, we delete all the small heterotachy subtrees from the original tree. For convenience, we will refer to such trees as large RAS/ER subtrees although technically failure to reject a null hypothesis does not imply its acceptance. It also should be noted that, in principle, corrections should be made for multiple testing in order that $P$ values have their usual meaning. The methods described thus should be considered exploratory rather than formal.
A Distance-Based Heterotachy Test

Because the ER model and the RAS model are nested in the general heterotachy model proposed by Wang et al. (2007), the “PROCOV test” (Wang et al. 2007) can also be used to distinguish between the general heterotachy model and the RAS model. For this test, the $P$ values should in principle be calculated taking into account the boundary conditions. Instead, because the general heterotachy model has four more parameter than the RAS model, the LRT uses a chi-square distribution with four degrees of freedom. The RAS model has one more parameter than the ER model, so a chi-square with 1 degree of freedom is used. Both these approaches give $P$ values that are larger than they should be. Though the computation of the PROCOV test is expensive, we will compare our results with those of the PROCOV test (Wang et al. 2007).

All the hypothesis tests are at the 5% level unless otherwise noted. Table 1 gives the $P$ values and conclusions drawn from $T_1$, $T_2$, and the PROCOV test. We will start with a simple example.

### Metazoan Nuclear Data

The metazoan nuclear data have 11 taxa and 50,462 amino acid sites and was originally considered in Dopazo H and Dopazo J (2005). The tree in figure 5 was obtained applying neighbor joining to PAHA-corrected distances and agrees with the tree estimated by Dopazo H and Dopazo J (2005) as well as the binned distance methods of Susko and Roger (2007). It should be noted, however, that upon removing fast-evolving sites, Dopazo H and Dopazo J (2005) found that Caenorhabitis elegans group with D. melanogaster and A. gambiae. This has been referred to as the Ecdysozoa tree and is more widely considered correct (see also Philippe et al. 2005). After analysis, we found that G1 was a small heterotachy subtree and G2 (see fig. 5) a large RAS subtree. There were no ER subtrees in this data set.

Compared with the results of the PROCOV test (see table 1), the PROCOV test fails to reject G2 as a RAS subtree with a $P$ value of 1.000, whereas the $P$ value for $T_2$ was 0.134. If we add two more taxa (D. melanogaster and A. gambiae) to G2, both $T_1$ and the PROCOV test reject the RAS model with $P$ value 0.000. Because the PROCOV test needs a tree with at least four taxa, the PROCOV test cannot be applied to subtree G1.

### Eukaryote Phylogenomic Data

We applied our tests to a large set of eukaryote phylogenomic data with 133 proteins from 40 taxa and 24,294 amino acid sites (Brinkmann et al. 2005). After analysis of

---

**Fig. 4.** Results of the simulation study showing the effect of the sequence length, shape parameter $\alpha$, the longest branch length, and the proportion of sites in the first partition on the level and the power of the within-gene heterotachy test $T_3$. (a) The shape parameter $\alpha$ was fixed at 0.25; (b) sequence length was fixed as 4,000 sites; (c) sequence length was fixed as 4,000 sites and the shape parameter $\alpha$ was fixed as 0.25; (d) sequence length was fixed as 4,000 sites and the shape parameter $\alpha$ was fixed as 0.25.
FIG. 5. The results of the tests $T_2$ and $T_1$ for the metazoan nuclear data. G1 is the smallest heterotachy subtree and G2 is the largest RAS subtree. We found no ER subtree for this data set.

this data set, we found that the only large ER subtree in this data set is G7 (see fig. 6). We also found that the only small general heterotachy subtree is G4. G3, G5, G6, G8, G9, and G10 were all classified as large RAS subtrees.

It is interesting that G7 was considered as an ER subtree in our tests, but the PROCOV test rejected it as an ER subtree with $P$ value 0.000 and failed to reject it as a RAS subtree with $P$ value 1.000. More interestingly, $T_2$ failed to reject G4 as a RAS subtree, where the PROCOV test rejected it as a RAS subtree with $P$ value 0.021. However, if we delete Rhodophyta from subtree G4, it will be considered as RAS subtree in both our test and the PROCOV test.

Subtree G8 also gave surprising results. From the $P$ value of $T_2$ and $T_1$, we cannot reject G8 as a RAS subtree, whereas the PROCOV test rejected it as a RAS subtree with $P$ value 0.000. However, if we add Glomus to G8, the PROCOV test could not reject the RAS hypothesis for G8 + Glomus with $P$ value 1.000. By contrast, $T_2$ rejects the RAS subtree with $P$ value 0.0088. If we add one more taxa Chytridiom, both our test and the PROCOV test reject the RAS subtree for G8 + Glomus + Chytridiom subtree with $P$ value 0.000. It appears that the PROCOV test is very sensitive to taxon selection.

Real Data Analysis for the Within-Gene Heterogeneity Test

We have implemented our test $T_3$ on the chloroplast genome sequences with 29 taxa, 57 genes, and 16,124 amino acid sites from Gruenheit et al. (2008). The heterotachous behavior of plastid rpo genes is well known (Lockhart et al. 1998, 2005; Ané et al. 2005). Most of the genes in this data set are <500 sites, which is problematic because test $T_3$ needs about 4,000 sites (the total number of sites for two genes, see fig. 4a) to get reliable results. Thus, we concatenated sequences according to the names of the genes (see table 3 in Gruenheit et al. 2008) and chose six groups:

- Group A: atpA, atpB, atpE, atpF, atpH;
- Group B: psaA, psaB, psaC, psaI;
- Group D: rpl14, rpl16, rpl2, rpl20;
- Group E: rpoB; rpoC1; rpoC2;
- Group F: rps11, rps12, rps14, rps18, rps19, rps2, rps3, rps4, rps7, rps8.

The first row in table 2 suggests that there is no within-gene heterotachy between Group B and Group C for subtree Groups 1, 2, 3, and 4 (subtree group designations refer to fig. 7). For this reason, we combined Group B and Group C together as one gene and then performed a within-gene heterotachy test with Groups A, E, and F.

From table 2, we find that the within-gene heterotachy test rejected the null hypothesis in Groups 3 and 4 for gene A with gene F, and also for gene Group B, C with gene Group E. We conjectured that gene E plays a major role in the rejection of the null hypothesis. In order to test this conjecture, we applied $T_3$ to each pair of the genes in Group E. The results are listed in the last three rows of table 2. We can see that $T_3$ strongly rejected the null hypothesis for each pair of the genes in Group 4, which means that there is within-gene heterotachy for each pair of genes in Group 4. Especially, for
Table 1. The \( P \) Values for the Tests \( T_2, T_1, \) and the PROCOV Test.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Taxa Group</th>
<th>Heterotachy test</th>
<th>Heterogeneity test</th>
<th>Conclusion of ( T_2, T_1 )</th>
<th>(PROCOV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>G1</td>
<td>0.000</td>
<td>NA</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>M</td>
<td>G2</td>
<td>0.134</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>E</td>
<td>G3</td>
<td>0.389</td>
<td>0.827</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>E</td>
<td>G4</td>
<td>0.000</td>
<td>NA</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>E</td>
<td>G5</td>
<td>0.054</td>
<td>NA</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>E</td>
<td>G6</td>
<td>0.173</td>
<td>0.021</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>E</td>
<td>G7</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>E</td>
<td>G8</td>
<td>0.033</td>
<td>0.000</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>E</td>
<td>G9</td>
<td>0.416</td>
<td>0.999</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>E</td>
<td>G10</td>
<td>0.056</td>
<td>0.183</td>
<td>0.000</td>
<td>NA</td>
</tr>
</tbody>
</table>

NOTE.—The first column indicates the data set under consideration. 'M' indicates the metazoan data and 'E' indicates the eukaryote data. The second column lists the taxa groups, which are indicated in figures 6 and 7. The third and the fourth columns list all the \( P \) values when the corresponding test is applied to the taxa groups. \( T_2 \) is the test for heterotachy, \( T_1 \) is the test for heterogeneity, and the PROCOV test is the LRT the general covarion model. The last column lists the conclusions drawn from \( T_2 \) and \( T_1 \). The conclusion for the PROCOV test was put brackets if it was different from our test. Otherwise, we marked as ‘*.’ ‘NA’ means not applicable because the PROCOV test needs a tree with at least four taxa. "GH" means general heterotachy model, "RAS" means rates-across-sites model, and "ER" means equal rates model.

Discussion

One issue of concern with the methods presented here is that there may not be much information in distances about both \( d \) and \( \alpha \). However, as shown in Wu and Susko (2010), if the fixed rate matrix used in the PAHA method has more than two distinct nonzero eigenvalues, \( \alpha \) and \( d \) will be identifiable. Because the rate matrix used in this article is the JTT rate matrix, which has 19 distinct eigenvalues, the parameters are identifiable. Wu (2010) also showed that when sequence length is larger than 4,000 sites, the mean square error of the estimation of \( \alpha \) and \( d \) becomes stable and low.

The simulations and the analysis of real data indicate that under a variety of conditions the heterogeneity test \((T_1)\) and the heterotachy test \((T_2)\) perform well. The level of the tests are below 5%, and the power of the tests increases with increasing sequence length. The tests have computational advantages and should work well with large sequence data sets. Figure 3c shows that when the longest branch length increases, the power of the test \( T_2 \) will increase as well. Note that increasing the longest branch length also increases the size of the tree, potentially affecting

![Fig. 6. The results of the tests \( T_2 \) and \( T_1 \) for the eukaryotic data. G4 is the smallest heterotachy subtree and G7 is the largest ER subtree. G3, G5, G6, G8, G9, and G10 are the largest RAS subtree.](image)
Table 2. The Results of $T_3$ and $T_2$ (in brackets) Applied to the Chloroplast Genome Sequences.

<table>
<thead>
<tr>
<th>Two Group of Genes</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>B with C</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>A with B, C</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>A with E</td>
<td>.</td>
<td>.</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>A with F</td>
<td>.</td>
<td>.</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>B, C with E</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>*</td>
</tr>
<tr>
<td>B, C with F</td>
<td>.</td>
<td>.</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>E with F</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>*</td>
</tr>
<tr>
<td>rpoB with rpoC1</td>
<td>.</td>
<td>.</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>rpoB with rpoC2</td>
<td>.</td>
<td>.</td>
<td>***</td>
<td>*</td>
</tr>
<tr>
<td>rpoC1 with rpoC2</td>
<td>.</td>
<td>.</td>
<td>***</td>
<td>**</td>
</tr>
</tbody>
</table>

NOTE.—A single asterisk, double asterisks, and triple asterisks indicate that the null hypothesis was rejected at $P = 0.05, 0.01,$ and 0.001 respectively. A dot indicates that the null hypothesis was not rejected. Groups 1–4 are indicated in figure 7.

The within-gene heterogeneity test ($T_3$) performs well under several conditions. The level of the test is close to 5%, and the power of the test increases with increasing sequence length. The power is low for smaller sequence lengths, however, suggesting the test should be used primarily with larger sequence length data sets. Simulation results also indicated that the test needs to be applied cautiously when $\alpha$ is small (see fig. 4b), where the power of $T_3$ is very low.

The value of the shape parameter $\alpha$ also affects the performance of the test $T_2$. If $\alpha$ increases, the null model approaches an ER model, but the model under the alternative hypothesis will still be a heterotachy model. Moreover, the level of the heterotachy will increase because when $\alpha$ approaches $+\infty$, the model under the alternative hypothesis will be dominated by heterotachy. Thus, the difference between null and alternative hypothesis in a sense becomes larger when $\alpha$ increases, which may be why the power of the test increases.

In order to show whether other types of evolutionary heterogeneity might cause the rejection of their non-heterotachy null hypothesis for the heterotachy test $T_2$, we simulated sequences under two cases through software INDELIBLE (Fletcher and Yang 2009). The simulation results (see Appendix A3 in Supplementary Material online) show that the heterotachy test $T_2$ will not reject heterogeneity under these two cases. In order to show the effect of the number of taxa on heterotachy test $T_2$, we also simulated sequences using six-taxon tree and eight-taxon tree. To evaluate the sensitivity of the heterotachy tests $T_2$, we varied several conditions just as same as we did for the four-taxon tree, except that the generating trees are changed. The simulation results (see Appendix A4 in Supplementary Material online) show that the number of taxa has little effect on the heterotachy test $T_2$.

When testing for heterotachy, it is assumed that $0 < \alpha < +\infty$. The usual likelihood theory applies for ML estimation of $\alpha$, which is used in justifying the use of a chi-squared distribution in $P$ value calculation. When testing heterogeneity, however, the generating $\beta = 0$. It is on the boundary of the parameter space that can create difficulties for likelihood theory (Self and Liang 1987). Careful study of the distributions obtained in Self and Liang (1987), which are for a somewhat different setting indicates that distributions tend to be skewed left of the distributions than would be used if boundary issues were ignored. This suggests that the power and level of the heterogeneity test might be less than it should be, which means that the heterogeneity test ($T_1$) is...
a conservative one. The results of our simulations were consistent with this conjecture.

Existing covarion models assume a time reversible process for rate evolution along the branches of the tree. Wu and Susko (2009) showed that the heterotachy model can be considered as multivariate rates-across-sites variation for pairs. Because this rates-across-sites distribution is implied by the same time reversible rate process, for each pair of taxa, the distribution might be the same. If this is true, we may not be able to distinguish a heterotachy model from a RAS model using pairs alone. It turns out that there are two cases of concern. If all pairs of taxa have the same evolutionary distance, the rates for these pairs will have the same distribution. In this case, the heterotachy test will fail to detect heterotachy model. However, if these pairs of taxa have different evolutionary distances, our heterotachy test will detect a heterotachy model successfully. To illustrate, recall that under the Tuffley and Steel (1998) covarion process, rates switch from OFF to ON and from OFF according to a stationary stochastic process. For two taxa separated by a short evolutionary distance, there is very little time for a switch from OFF to ON or ON to OFF. The implied rates are thus more likely to be large (the process is always ON) or small (the process is always OFF) than when the evolutionary distance is large. In the latter case, more intermediate rates are expected as the process is expected to switch more frequently.

Test $T_1$ and $T_2$ can detect whether a tree (subtree) is a heterotachy tree, RAS tree, or an ER tree. Compared with the tests proposed by Lockhart et al. (1998) and Ané et al. (2005), our tests do not depend on tree topology (though we use an assumed tree topology to find the ER and RAS subtrees), and the tree does not need to be split into several groups. Moreover, their tests use the information in the sequences only through counts of site patterns where some subsets of taxa are invariable. However, our test cannot detect the sites that cause heterotachy. One possible way to do this is by deleting conspicuous sites and applying the heterotachy test $T_2$. If the null hypothesis cannot be rejected, it suggests these sites might be heterotachous sites. One could also use a jack-knifed test statistic to give a relative ranking of the sites.

In the real data analysis, we found that our tests can efficiently detect general heterotachy (except in G6 and G8). Compared with the PROCOV test, our test gave the same results as those of the PROCOV test most of the time. Because the PROCOV test needs a tree with at least four taxa, we cannot compare our test with the PROCOV test when there are three taxa. Applying $T_1$ to the real data, we found two three-taxa ER subtree and one four-taxa ER subtree. As we can see from the table 1, $T_2$ is competitive to the PROCOV test and has the advantage of being less computationally expensive than the PROCOV test.

Because of functional constraints, the ER model is not realistic for real data. In our experience, if a data set has many taxa, the general heterotachy model is usually the only choice. However, one can identify subtrees that seem to be homogeneous in their rate variation using the ideas illustrated in our analysis of real data. Perhaps, not surprisingly the groups found tended to be separated from each other by relatively long branches.

In some analyses, $T_1$ gave significant result but $T_2$ did not. In other words, we concluded that there is no heterotachy in a data set after applying $T_2$ but concluded that there is within-gene heterotachy in this data set after applying $T_3$. This might be caused by the partition of sites in genes. In figures 3b and 4d, if the proportion of sites in partition (group of genes) 1 decreases from 0.5 to 0.1, the power of test $T_2$ decreases but for $T_3$ it increases. If two partitions (group of genes) are unbalanced (one is close to 0%, and the other is close to 100%), $T_3$ will tend to reject the null hypothesis, whereas $T_2$ tend to not reject. This might cause this behavior, though we do not know the real proportion of sites in partitions (group of genes) for a real data set. We assume a conventional general time reversible model. In practice, we have used the JTT model. Evolution is assumed independent across sites. Conditional on the amino acid at their internal node, evolution is independent along adjacent edges. Finally, evolution along an edge is according to a time reversible constant time Markov model with a rate matrix that is constant throughout the tree (and known).

**Supplementary Material**

Appendices A1–A4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We wish to thank Dr. A.J. Roger, Dr. H. Philippe and 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.

**References**


Gruenheit N, Lockhart PJ, Steel M, Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the


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


