Fast Statistical Tests for Detecting Heterotachy in Protein Evolution

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

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

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

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

***Corresponding author:** E-mail: hcwang@mathstat.dal.ca.

Associate editor: Barbara Holland

Abstract

The w statistic introduced by Lockhart et al. (1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol. 15:1183–1188) is a simple and easily calculated statistic intended to detect heterotachy by comparing amino acid substitution patterns between two monophyletic groups of protein sequences. It is defined as the difference between the fraction of varied sites in both groups and the fraction of varied sites in each group. The w test has been used to distinguish a covarion process from equal rates and rates variation across sites processes. Using simulation we show that the w test is effective for small data sets and for data sets that have low substitution rates in the groups but can have difficulties when these conditions are not met. Using site entropy as a measure of variability of a sequence site, we modify the w statistic to a w' statistic by assigning as varied in one group those sites that are actually varied in both groups but have a large entropy difference. We show that the w' test has more power to detect two kinds of heterotachy processes (covarion and bivariate rate shifts) in large and variable data. We also show that a test of Pearson's correlation of the site entropies between two monophyletic groups can be used to detect heterotachy and has more power than the w' test. Furthermore, we demonstrate that there are settings where the correlation test as well as w and w' tests do not detect heterotachy signals in data simulated under a branch length mixture model. In such cases, it is sometimes possible to detect heterotachy through subselection of appropriate taxa. Finally, we discuss the abilities of the three statistical tests to detect a fourth mode of heterotachy: lineage-specific changes in proportion of variable sites.

Key words: covarion, heterotachy, parametric bootstrap, entropy, phylogenetics.

Introduction

Given an alignment of protein sequences from two monophyletic groups of taxa, five types of site patterns are readily identifiable by visual inspection (Lockhart et al. 1998): type 1 sites have the same residue in both groups; type 2 sites have different residues between the groups but have the same residue within the same group; types 3 and 4 sites have variable residues in one group but the same residue in the second group; and type 5 sites have variable residues in both groups. Types 3 and 4 sites are widely recognized as having a typical "covarion" site pattern: sites are variable in one clade but invariable in the other clades or vice versa (Fitch and Markowitz 1970). This feature of changing the rate of variation at sites in different sequences is more generally called "heterotachy" meaning "different speeds" of evolution (Lopez et al. 2002). Based on the differences in temporal and spatial distributions of rate shifts, several heterotachy models have been developed, including (see fig. 1) 1) the covarion models where rate change is a gradual and stochastic process acting on all branches of a phylogeny (Tuffley and Steel 1998; Galtier 2001; Huelsenbeck 2002; Wang et al. 2007; Whelan 2008), 2) bivariate rate shift models where sudden changes in rates at multiple sites occur at a particular split in the tree and the two subtrees undergo independent rate variation across sites (Susko et al. 2002), 3) mixture of branch length models where a proportion of sites are generated from a tree having one set of branch lengths and the remaining sites generated from the same tree but having a different set of branch lengths (Kolaczkowski and Thornton 2004, 2008; Spencer et al. 2005; Zhou et al. 2007; Pagel and Meade 2008), and 4) lineage-specific variation in the proportion of variable sites (Lockhart et al. 1996, 2006; Lopez et al. 2002). These various forms of heterotachy can generally be viewed as a multivariate distribution of rates-across-sites variations (Wu and Susko 2009). Various simulation and empirical studies have shown that ignoring the heterotachy property of sequence evolution may lead to topological biases that will mislead tree building (e.g., Lockhart et al. 1996, 2006; Kolaczkowski and Thornton 2004; Wang, Susko et al. 2008). Furthermore, analyzing heterotachy properties can be used to detect functional divergence of protein families (e.g., Gu 1999; Gaucher et al. 2001; Knudsen and Miyamoto 2001; Pupko and Galtier 2002; Penn et al. 2008; Wang et al. 2009; Studer and Robinson-Rechavi 2010) and positive selection (Siltberg and Liberles 2002; Guindon et al. 2004; Dorman 2007).

Because of the importance of heterotachy processes in protein evolution and phylogenetic studies, computational methods have been developed to detect whether heterotachy has played a role in the evolution of a protein family

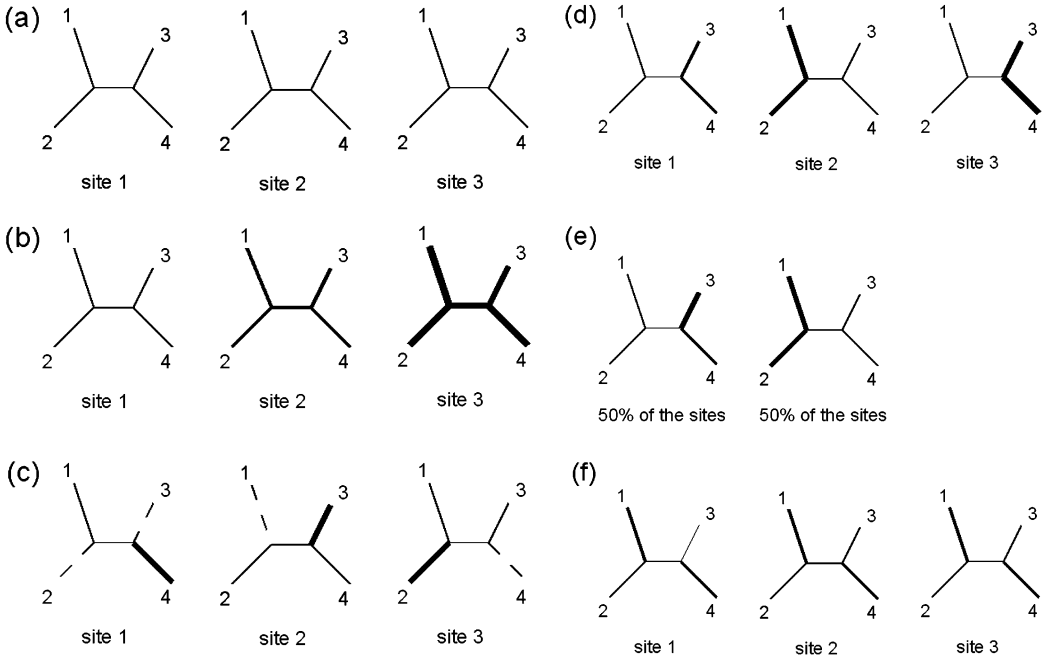

FIG. 1. Various models of rate variation across sites and lineages. The thickness of the lines represents different rates: the thicker the line, the higher the rate is at the site for the lineage. (*a*) equal rates; (*b*) rates-across-sites; (*c*) covarion; (*d*) bivariate rates—under this model the two subtrees are allowed to have different rates and the rates vary across sites independently between the two subtrees; (*e*) branch length mixture; (*f*) lineage-specific variation in proportions of variable sites. (*c*)–(*f*) are heterotachy models. The solid and dashed lines in (*c*) represent that the lineage can be switched "ON" or "OFF" for substitution, respectively. This figure is partially adapted from Wu and Susko (2009).

(e.g., Lopez et al. 1999; Baele et al. 2006). Most of the methods employ likelihood ratio tests (Gaucher et al. 2001; Knudsen and Miyamoto 2001; Pupko and Galtier 2002; Penn et al. 2008; Wang et al. 2009) to compare the relative fit of a heterotachy model versus a default homotachy model, which is typically a rates variation across sites (RAS) mixture model based on a discretized gamma distribution of site rates. These methods, although statistically sound, have the disadvantage of being computationally intensive (Wang et al. 2009).

Heterotachy, whatever form it takes, will ultimately show heterotachous patterns at the sites of the sequence alignment. For instance, a protein family evolving according to a covarion process will necessarily have higher proportions of types 3 and 4 sites in the sequences than a protein family evolving in a homotachous mode. Sequences simulated under different models including the equal rates (ER), RAS, and a general covarion model (COV) (Wang et al. 2007) for the same tree (an elongation factor [EF] tree from an alignment of 13 bacterial elongation factor Tu [EF-Tu] and 17 elongation factor 1α [EF-1α] sequences) show that the numbers of the five site types are very different among these models: ER has the highest number of type 5 sites and very few sites of types 1 and 2; RAS has fewer type 5 sites than under the ER model but its number is greater than types 3 and 4 sites combined; COV has more types 3 and 4 sites combined than type 5 sites; both RAS and COV have similar numbers of types 1 and 2 sites. Therefore, in principle, the different models of evolution may be distinguished by comparing the numbers of site types they induce in the data. Indeed, the elegant *w* statistic, that compares the proportions of the types 3 and 4 sites and type 5 sites in an alignment, was first proposed by Lockhart et al. (1998) to test whether covarion or RAS processes generated the data under consideration. Ané et al. (2005) further improved the *w* test by obtaining null distributions for hypothesis testing using a parametric bootstrapping approach. Other studies have applied the test to phylogenetic inference (Lockhart et al. 2000) and empirical data for detecting covarion evolution (Gruenheit et al. 2008).

Let N_i denote the numbers of type *i* sites ($i = 1, 2, 3, 4, 5$) and N is the total number of sites in an alignment. The *w* statistic originally given in Lockhart et al. (1998) and modified in Ané et al. (2005) is

$$w = \frac{N_5}{N} - \frac{(N_3 + N_5)(N_4 + N_5)}{N^2}. \quad (1)$$

The first term on the right-hand side of the equation is the proportion of the sites varying in both groups (N_5 sites), and the second term is the product of the proportions of sites varying in one group and that in the other group. Under the ER model, as there is no correlation between the variabilities of the two groups of taxa, the two terms should be approximately equal so that the expected value of *w* is 0. Under the RAS model, if one group of taxa is variable at a particular site, then other groups of taxa are more likely to be variable at that site. Therefore, the RAS model induces a strong correlation between the variability of one group and other groups at sites. The N_3 and N_4 should be small so that *w* is positive. Under the

COV models, there is less correlation between the variabilities of the two groups and so N_3 and N_4 will be large. Therefore, although w is expected to be positive, it should be smaller under COV than under the RAS model (Ané et al. 2005). Although the foregoing definitions are reasonable, it should be noted that the N_3 and N_4 sites only represent a subset of patterns expected under the covarion model. For example, covarion sites may in fact display variable residues in both groups but one group may be more variable than the other (Wang et al. 2009). The above expression for w effectively ignores these sites, and this may, we suspect, be part of the reason why w is quite sensitive to taxon sampling (Gruenheit et al. 2008). In the simulation settings we will consider, given a large alignment with many taxa in each group, it is possible that there are always at least a few variant residues at each site in both groups. Under these conditions, the w test should have very low power to detect covarion evolution or other heterotachy processes. We propose a modified w test and correlation test designed to detect heterotachy in such cases.

Methods

A Modified w Statistic (w') for Covarion Tests

The w statistic effectively checks whether there are more sites than expected under RAS that are variable in one group but not the other. Such sites likely have higher rates in the variable group and lower rates in the other. Similarly, evidence for heterotachy is also provided by an excess of sites that vary in both groups but that show much less variability in one of the groups. We modify the w test by adding such sites to the counts of those that vary in only one of the groups. Let

$$w' = \frac{N_5 - N_3' - N_4'}{N} - \frac{(N_3+N_3'+N_5-N_3'-N_4')(N_4+N_4'+N_5-N_3'-N_4')}{N^2} = \frac{N_5 - N_3' - N_4'}{N} - \frac{(N_3+N_5-N_4')(N_4+N_5-N_3')}{N^2}, \quad (2)$$

where N_3, N_4, N_5, and N have the same meaning as before and N_3' (respectively, N_4') is the number of the N_5 sites that have a large rate difference that warrant them to be considered to be N_3 or N_4 sites. The rate, or the variability of a site, can be measured by the Shannon entropy, which is calculated as follows:

$$H = -\sum_{j=1}^{20} p_j \times \log(p_j), \quad (3)$$

where p_j is the amino acid frequency vector at the site. Higher entropy means the site is highly variable. The values range from 0, when all sequences have the same amino acid at a site, to a maximum of 4.3219, when all 20 amino acids are represented and equally frequent at the site. A covarion site of the N_3 or N_4 type will have zero entropy in one group and nonzero entropy in another group, and consequently, this site will usually have a big entropy difference between the two groups when the variable group has several different amino acids at the site.

To determine the critical value of entropy difference that assigns an N_5 site to be N_3' or N_4' site, we simulate a data set of 100,000 sites under an RAS model for the same tree estimated from the target data and calculate the 5% percentile of the distribution of the entropy difference of the 100,000 RAS sites. If the entropy difference of an N_5 site in a target sequence alignment is bigger than this 5% percentile value, then it will be assigned to an N_3' or N_4' site.

A Site Entropy-Based Correlation Test

Under an RAS model, site variation in one group of sequences is positively correlated with the variation in the other group of sequences. Under heterotachy models like the covarion model, rate variation at sites within each of the two groups is also correlated because the variability of rates changes gradually along the tree. However, for the covarion model the correlation will be weaker than for RAS as sites switching from ON to OFF and from OFF to ON diminishes the correlation. Under an ER model, the variability of one group is not correlated with that of the other group. Therefore, if we measure site variability by the Shannon entropy of the site, we can distinguish between the covarion and RAS models as a weaker correlation between site entropies under covarions than under RAS. Given two sets of N site entropies calculated as in equation (3) for clades 1 and 2, the Pearson correlation coefficient r between site entropies of the two clades is computed using the equation

$$r = \frac{\sum_i^N (H_1^{(i)} - \bar{H}_1)(H_2^{(i)} - \bar{H}_2)}{\sqrt{\sum_i^N (H_1^{(i)} - \bar{H}_1)^2 \sum_i^N (H_2^{(i)} - \bar{H}_2)^2}}, \quad (4)$$

where $H_1^{(i)}$ and $H_2^{(i)}$ are the entropies for clades 1 and 2 at site i and $\bar{H}_1$ and $\bar{H}_2$ are the average entropy across the sites for clades 1 and 2, respectively.

For the three tests (w, w', and the Pearson correlation coefficient of site entropy), we used parametric bootstrapping (Ané et al. 2005) to obtain null distributions of the test statistics under the RAS model. If the test statistic is within the left 5% tail of the null RAS distribution, then the RAS model is rejected at the 5% level. The three covarion tests, partially modified from the code for calculating the w statistic (Ané et al. 2005), are implemented in covTests.c (http://www.mathstat.dal.ca/~hcwang/Procov/).

In addition to testing for covarion processes, we also applied the three tests to data simulated under three other heterotachy models: an uncorrelated bivariate rate model where rates are independently assigned between two monophyletic groups (Susko et al. 2002; see fig. 1*d*); a branch length mixture model where there are multiple sets of branch lengths for the same topology for different site partitions of the alignment (Kolaczkowski and Thornton 2004; see fig. 1*e*); and a model that allows lineage-specific changes in proportion of variable sites (Lockhart et al. 1996; Shavit Grievink et al. 2008; see fig. 1*f*).

Simulations

As different settings are used for simulating different heterotachy models, we summarize the models, trees, parameter settings, and simulation programs as well as null

Table 1. Models, Parameters, and Programs for Simulating Heterotachy.

Model	Example Tree	Sequence Length	Parameters (number of simulated data sets = 100)	Program	Null Distribution (for each heterotachy data set, simulate 1,000 RAS data sets)
General covarion model	Figure 2	1,000	WAG + Γ (α), α = 0.5; s01 = 0.5; s10 = 0.5; s11 = 0.5; π = 0.75	Seq-gen-aminocov	WAG + Γ (α); 1,000 sites
Bivariate model	Figure 2	1,000	WAG + Γ (α), α = 1.0; two subsets were simulated separately from the same root sequence	indel-seq-gen	WAG + Γ (α); 1,000 sites
Branch length mixture model	Figure 3	1,000	WAG + Γ (α), α = 0.5; half of the sites from the left or right trees in figure 3	Seq-gen	WAG + Γ (α); 1,000 sites
Branch length mixture model with taxon subselection	Figure 4	1,000	WAG + Γ (α), α = 0.5; half of the sites from the left or right trees in figure 4	Seq-gen	WAG + Γ (α); 1,000 sites
Lineage-specific change in p_{var}	Figure 2 in Shavit Grievink et al. (2008)	10,000	WAG + Γ (α) + I + event, α = 0.5, p_{inv} = 0.8; event p_{var} = 0.2	LineageSpecificSeqGen	WAG + Γ (α) + I; 10,000 sites

distributions in table 1 and describe them in detail in the following subsections.

Covarion Simulations

The first tree used in our simulations is shown in figure 2. This topology has two monophyletic groups each with the same number of taxa and each forming a subtree in the shape of a star tree connected to the other subtree by an internal edge, which has length *b*. By using the star tree for the two subtrees, we can minimize the number of settings for edge lengths within the subtrees so that the simulation will be targeted to overall edge lengths. Each edge within the two subtrees has an identical length of *a*. In all simulations, we set the internal edge length to $b = 1.0$ substitution per site and varied edge lengths *a* to be 0.1, 0.3, 0.5, or 0.7. To look at the effect of the size of the subtrees on the analyses, we varied the numbers of taxa per clade to be 5, 10, 15, and 20, respectively, for different simulations. Furthermore, in order to check the efficiency of the methods we also simulated very large data sets (150 taxa per clade) with 1,000 amino acid sites and varied edge lengths *a* to be 0.1, 0.3, 0.5, or 0.7. Computational times for this size of data with some likelihood-based methods, such as using PROCOV (Wang et al. 2009), are prohibitive.

We considered both the type I error and the power of the tests. The type I error rate is the probability of rejecting the null hypothesis when it is true. For 100 simulations, the type I error for a 0.05-level test should be within $0.05 \pm 1.96\sqrt{\frac{0.05\times0.95}{100}}$ (i.e., in the range of 0.73–9.27%). The power of a test is the probability of correctly rejecting the null hypothesis. To check the type I error of the tests, we simulated 100 data sets of 1,000 sites under the WAG + Γ (α) model (set $\alpha = 0.5$) using seq-gen (Rambaut and Grassly 1997). To evaluate the power of the tests, we simulated 100 data sets under the general covarion model, using seq-gen-aminocov (Wang, Li et al. 2008) to generate sequences of 1,000 sites under the WAG + Γ (α) model (set $\alpha = 0.5$) and covarion parameters s01 = 0.5; s10 = 0.5; s11 = 0.5 and covarion "switch-on" frequency $\pi = 0.75$ (these parameter settings represent the average estimated covarion parameters for most of the empirical data sets we tested; see Wang et al. 2007). We then used RAxML (Stamatakis 2006) under the WAG + Γ (α) model to estimate α from each of the simulated 100 data sets by fixing the tree as the simulating tree but optimizing edge lengths. We used each of the 100 estimated α values and trees to simulate, separately, 1,000 data sets of the same number of taxa and 1,000 sites under the WAG + Γ (α) model with seq-gen. We calculated the three quantities (*w*, *w*′, and the Pearson correlation coefficient) for the RAS- or COV-simulated data sets and each of the corresponding 1,000 RAS-simulated data sets. For the type I error evaluation, if any of the three quantities for the RAS-simulated data is less than the 50th smallest corresponding quantity among the 1,000 RAS-simulated replicated data sets, then that quantity incorrectly rejects the RAS process in the data; otherwise, it correctly recognizes it to be an RAS data set. For the power evaluations, if any of the three quantities for the COV-simulated data is less than the 50th smallest corresponding quantity among the 1,000 RAS-simulated data sets, then that quantity correctly assigns the COV-simulated data as a covarion data set; otherwise, it wrongly recognizes it to be an RAS data set.

Simulations under a Bivariate Rate Model

To evaluate the powers of the *w*, *w*′, and correlation tests for data simulated under the bivariate rate model, we used indel-seq-gen (Strope et al. 2009) under a WAG + Γ (α)

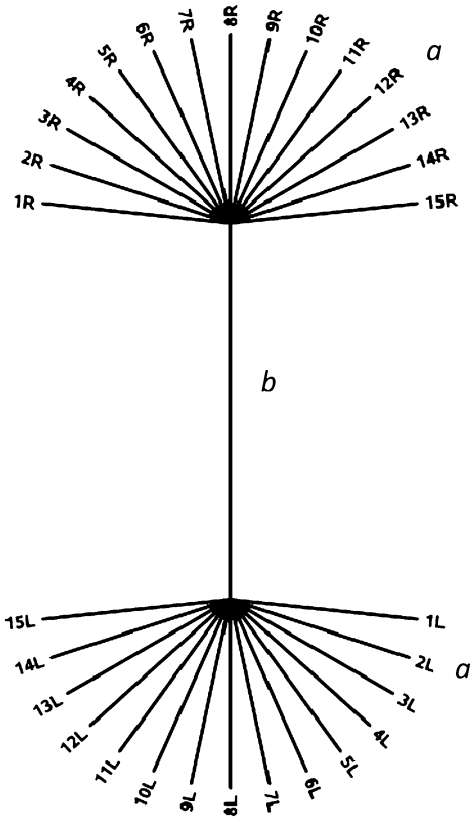

FIG. 2. One of the four simulating trees with two "star" subtrees separated by an internal edge. Here a tree of 15 taxa per group is shown. Other simulating trees have 5, 10, and 20 taxa per group. *b* is the length of the central edge connecting the two subtrees, which is set to 1.0; *a* is the length of the edges within the two subtrees, which is set to 0.1, 0.3, 0.5, and 0.7, respectively, for different simulations.

model to get a root sequence of 1,000 amino acid sites for the simulating tree (fig. 2). We set α to 1.0 in this simulation as prior simulations indicated that all the tests will have a perfect power when setting $\alpha = 0.5$. Using the generated root sequence, we used indel-seq-gen under a WAG + Γ (α) model (set $\alpha = 1.0$) to independently simulate sequences of 1,000 sites based on the upper and lower subtrees in figure 2. We then combined the two sequence data sets from the two subtrees into one data set. Because indel-seq-gen randomly assigns sites to different rates according to the gamma distribution, the end result is that the two independently simulated subtrees are evolved from the same root sequence but display a bivariate rate shift at sites between the subtrees. We repeated this procedure 100 times to generate 100 sets of bivariate rate data. For each of these data sets, we used RAxML under the WAG + Γ (α) model to estimate an α by fixing the tree as shown in figure 2 while optimizing edge lengths. We then used the estimated α and corresponding estimated trees to simulate 1,000 data sets of 1,000 sites under the WAG + Γ (α) model with seq-gen. The three test statistics were calculated for each of the bivariate data sets, and their RAS replicates and the power of the tests were determined similarly as in the covarion test.

Simulations under a Branch Length Mixture Model

To evaluate the powers of the three tests on data simulated under a branch length mixture model, we used seq-gen to simulate 100 data sets of 1,000 sites each under WAG + Γ (α) model (set $\alpha = 0.5$) with two equal partitions of sites that have the same tree topology but different sets of edge lengths for the left and right trees (fig. 3). We simulated data using trees with 8, 16, 24, and 32 taxa, respectively. For each data set, we used RAxML under the WAG + Γ (α) model to estimate α using a tree topology as shown in figure 2 as fixed topology while optimizing the branch lengths. We then used the estimated α and corresponding estimated tree to simulate 1,000 data sets of 1,000 sites under the WAG + Γ (α) model. The three test statistics were calculated for each of the branch length mixture data, and their RAS replicates and the significance of the tests were determined similarly as in the covarion test.

Furthermore, for the trees harboring 8 and 16 taxa, we subselected a taxon in each clade to have a different edge length than the rest of the taxa in the same clade. Figure 4 shows a pair of subselected trees of eight taxa that are different from the balanced tree shown in figure 3. We used the same procedure to apply the three tests on data simulated under this kind of trees.

Simulations under Lineage-Specific Variation in Proportion of Variable Sites (p_{var})

To investigate whether the *w*, *w*′, and correlation tests can be used to detect this type of heterotachy, we used LineageSpecificSeqGen (Shavit Grievink et al. 2008) to simulate

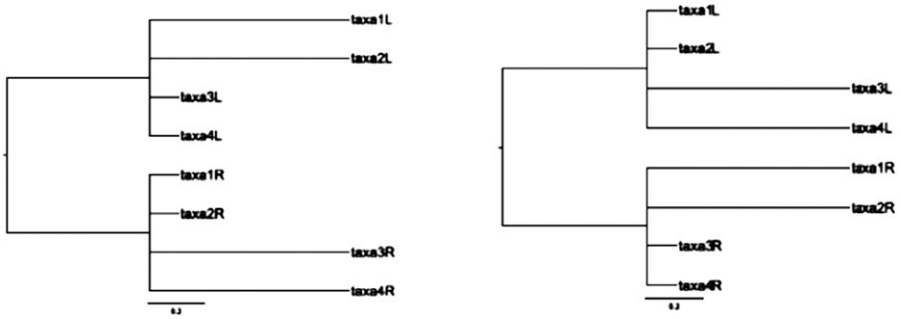

FIG. 3. One pair of the simulating trees for a mixed branch length model. Here a tree of four taxa per group is shown. Other simulating trees have 8, 12, and 16 taxa per group. The length of the central edge connecting the two groups is set to 1.0. The short and long edges in the two groups have a length of 0.1 and 0.7, respectively.

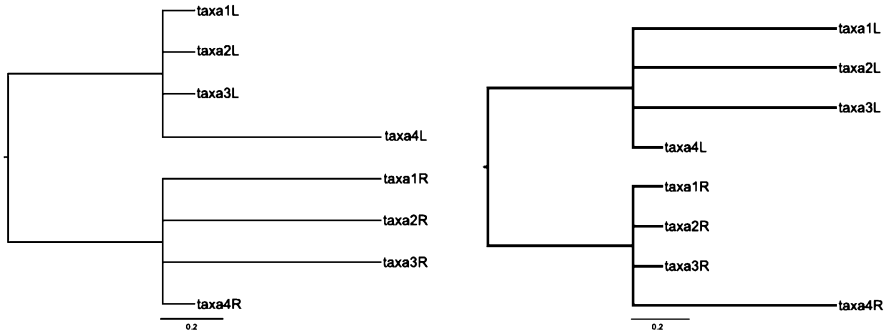

Fig. 4. Subselecting taxa for the mixed branch length model. Here a pair of eight-taxon trees is shown in which one taxon has a different edge length than the other three taxa in the same clade. In another simulation setting, a 16-taxon tree is subselected so that 1 taxon has different edge length than the other 7 taxa in the same clade. The length of the central edge connecting the two groups is set to 1.0. The short and long edges within the two groups have a length of 0.1 and 0.7, respectively.

100 data sets of 16 taxa each in which two nonsister lineages have correlated changes in their p_{var} (the tree is shown in fig. 2 of Shavit Grievink et al. 2008). As specified in Shavit Grievink et al. (2008), we simulated 10,000 sites for each data set setting the two events $p_{var} = 0.2$ and the proportion of invariable sites at the root node of the tree (p_{inv}) = 0.8, WAG + Γ ($\alpha = 0.5$) using four rate categories. For each data set, we used RAxML to estimate a tree under WAG + Γ (α) + I model with four rate categories by fixing the topology as the simulating tree. Using each of the 100 estimated α's, p_{inv}'s, and corresponding trees, we simulated 1,000 data sets of 10,000 sites under the WAG + Γ (α) + I model. The three test statistics were calculated for each of the p_{var} change data, and their RAS replicates and the significance of the tests were determined similarly as in the covarion test.

Four Empirical EF-Related Data Sets

To investigate the performance of the w, w', and correlation tests on real data, we considered four previously studied sequence data sets of EFs. The first data set consists of orthologous sequences of a 17-taxon subtree of eukaryotic elongation factor 1α (eEF-1α) and a 13-taxon subtree of bacterial EF-Tu and has 380 aligned sites (Gaucher et al. 2001). The second data set of EF-1α has 349 sites and 24 taxa including 2 archaeal taxa and 22 eukaryotic taxa, 3 of which are Microsporidia (Inagaki et al. 2004). The third data set consists of orthologous sequences of a 27-taxon subtree of eEF-1α and a 13-taxon subtree of archaeal EF-1α (aEF-1α). The fourth data set is a data set of paralogous sequences consisting of a 13-taxon subtree of Hsp70 subfamily B suppressor 1 (HBS1) and a 17-taxon subtree of eukaryotic release factor (eRF3). The third and fourth data sets have 269 sites (Susko et al. 2002). For each EF-related data set, we used RAxML to infer a phylogenetic tree under the WAG + Γ (α) model. We then used the estimated tree and the α shape parameter to simulate 1,000 data sets with the same sequence length as the original EF-related data under the RAS model. For the second data set (Microsporidia EF data), a Microsporidia–Archaea clade tree, as expected, would be estimated under the gamma model, which was slightly different from the true tree that unites the Microsporidia with fungi. We used both the wrong and the true trees for simulating the 1,000 replicates of the RAS data. The w, w', and Pearson correlation coefficients were calculated for each of the original data sets, and corresponding 1,000 RAS replicates and the significance of the tests were determined similarly as above. The four data sets are available at http://www.mathstat.dal.ca/~hcwang/Procov/.

Results

Performance of the Covarion Test in Simulation

Figure 5A–C shows the type I error rates of the w, w', and the correlation tests for data simulated under the RAS model and the power of the tests to distinguish the general covarion model from an RAS model. The type I error rates for all tests are close to the target of 5% (the α level) regardless of the number of taxa per clade and the edge lengths within the subtrees. Keeping the internal edge connecting the two subtrees b the same ($b = 1.0$), for very small edge lengths within the subtrees ($a = 0.1, 0.3$), w had good power to distinguish COV-simulated data from RAS for small and large trees. However, when a was increased to 0.5 and 0.7, w had a good power only when the simulating tree was small (i.e., 5 and 10 taxa per clade) and had very low power when the simulating tree had 15 or 20 taxa per clade. In contrast, both the w' and the Pearson coefficient had very good power regardless of the number of taxa and edge length a. The correlation test essentially had 100% power for the data simulated under the various settings.

For the analyses of the simulated very large data set (150 taxa per clade and 1,000 sites) under the general covarion model, on average it took 2.5 min to complete an analysis on a 3-GHz Dual Xeon E5450 with 16 GB RAM, which included the bootstrapping of 1,000 simulated RAS data sets of the same size. The results showed that both the w and the w' tests had reduced power to reject RAS. The w test had a power of 11% for the edge length $a = 0.1$, and it was further reduced to 0% and 2% for $a = 0.3$ and 0.5. Compared with the w tests the power of the w' tests was higher, being 65%, 5%, and 49% for $a = 0.1$, 0.3, and 0.5,

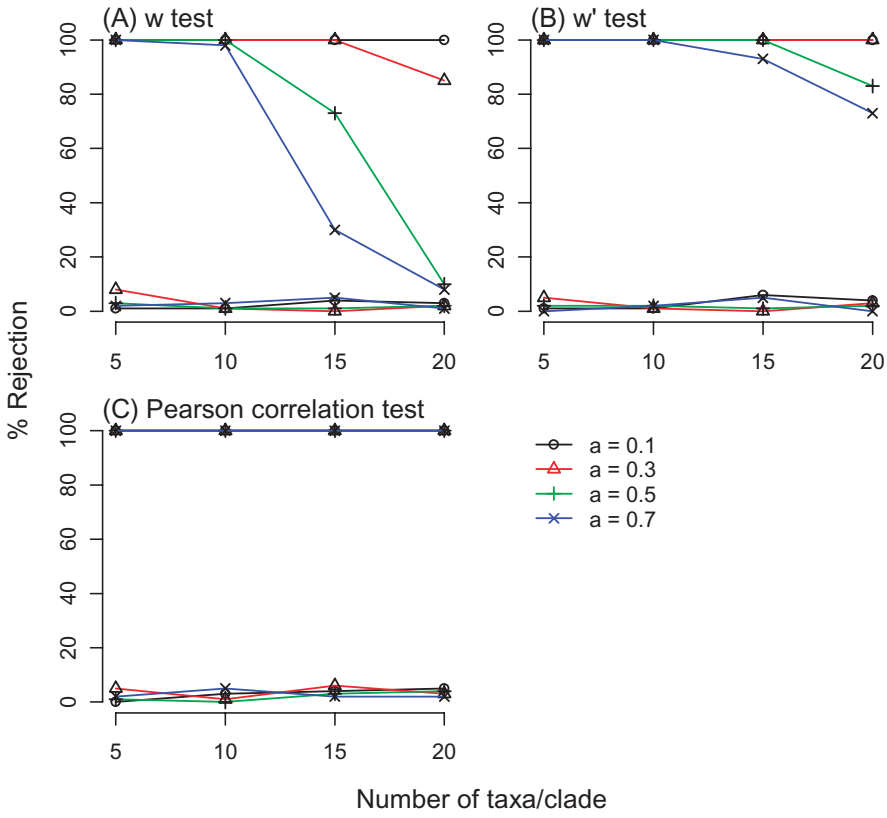

FIG. 5. The type I error rates of the three test statistics for data simulated under the RAS model (lower panel of lines) and the power of the tests for data simulated under a general covarion model (upper panel of lines). (*A*) *w* test; (*B*) *w′* test; (*C*) the correlation test. The simulating tree (see fig. 2) has 5, 10, 15, and 20 taxa per clade and a terminal edge length $a = 0.1$, 0.3, 0.5, and 0.7, respectively. The power of correlation tests is always 100% for different sizes of data, merging all lines into a single line (blue line in *C*). (*B*) Also shows some overlapping of the lines due to the same power for different size of the data.

respectively. However, both the *w* and the *w′* tests had elevated power (38% and 84%, respectively) for $a = 0.7$. This may seem surprising as with smaller numbers of taxa, power generally decreased with increasing *a*. The reason for the elevated power here is that for a very large number of taxa and large edge lengths within the groups, almost all sites become variable in the two groups (i.e., type 5 sites) and there are very few type 3 or 4 sites. Indeed, the distribution of the *w* values for data simulated under the RAS model is within a narrow range close to 0. Some of the *w* values for data simulated under the COV model become negative. These combined effects led the *w* test to reject RAS and thus have an unexpectedly higher power for the $a = 0.7$ cases. The *w′* test had a similar problem; the values of *w′* became negative for the COV- and RAS-simulated data leading to the rejection of RAS. Despite these difficulties associated with the *w* and *w′* tests when applied to the very large data sets, the correlation test, however, remained highly capable with its power reaching 100% for the 300-taxon data and all edge lengths used in the simulations.

Performance on Data Simulated under the Bivariate Rate Model

For data simulated under the bivariate rate model, both *w′* and the correlation tests had perfect power regardless of the number of taxa and the edge length *a*. The *w* test for the bivariate data (fig. 6), in general, had better power than for the covarion simulated data. It had perfect power for smaller edge lengths within the subtrees ($a = 0.1$, 0.3) and smaller trees (less than 15 taxa per clade). For the 20 taxa per clade tree, its power was reduced to 83% when the edge length *a* was 0.7, but this is still much better than the power observed for the COV-simulated data (comparing with fig. 5).

Performance on Data Simulated under the Mixed Branch Length Model

For the data simulated under the tree shown in figure 3, none of the three tests (*w*, *w′*, and the Pearson correlation) had any power to reject RAS regardless of the number of taxa used in the simulation. For this result, we used the α estimated from the simulated mixed branch length data to generate a null RAS distribution. We also separately estimated an α from each of the two subtrees for the simulated data and then averaged them to get an average α. Using this average α to simulate the RAS distribution, there was still no power to reject the null RAS distribution for all the three tests on this kind of data.

For the mixed branch length data with taxa subselection (see Methods), the power of rejecting RAS did increase and was dependent on the size of the simulating

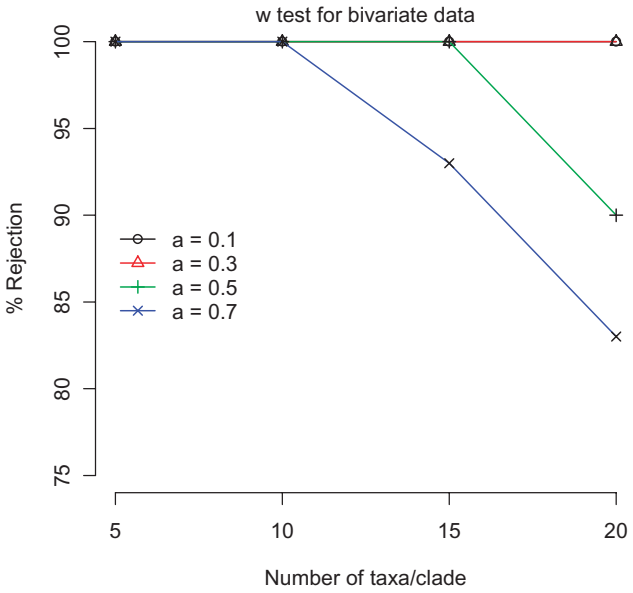

FIG. 6. Power of the *w* tests for data simulated under the bivariate rate model. The gamma shape parameter $\alpha = 1.0$ was used in the simulations. The simulating tree (see fig. 2) has 5, 10, 15, and 20 taxa per clade and a terminal edge length $a = 0.1$, 0.3, 0.5, and 0.7, respectively. Because the power for the 10 and 15 taxa per clade trees are all same (100%) for $a = 0.1$ and $a = 0.3$, the black line is overlapped with the red line and becomes invisible.

tree. For the eight-taxon trees with subselected taxa (fig. 4), the *w*, *w*′, and correlation tests had a power of 2%, 26%, and 88%, respectively. For the 16-taxon trees with subselected taxa, the power of the *w* test increased to 82%, whereas both the *w*′ and the correlation tests reached a perfect power.

Performance on Data Simulated under Lineage-Specific Variation in p_{var}

We used LineageSpecificSeqGen to simulate 100 data sets of 16 taxa based on the tree as shown in figure 2 of Shavit Grievink et al. (2008), in which two non-sister lineages have correlated changes in their p_{var}. Comparing the *w*, *w*′, and the Pearson coefficients for these data sets and corresponding WAG + Γ (α) + I replicated data sets (without p_{var} change), all three tests always rejected RAS + I for the changed p_{var} data sets. As a control, for data sets simulated under a WAG + Γ (α) + I model, which only lacked the heterotachy component of the changed p_{var} events, the three tests were not able to reject RAS + I in 99, 98, and 92 of 100 cases, respectively.

Empirical Data: The Four EF-Related Data Sets

A previous study based on examining rate difference between the eEF-1α subtree and bacterial EF-Tu subtree indicated that the EF-1α/EF-Tu data set evolved under a covarion-like process (Gaucher et al. 2001). This was further verified by direct covarion-based modeling with PROCOV (Wang et al. 2009). Here we applied the *w*, *w*′, and correlation tests to the data and found that all tests rejected RAS with high probability (table 2).

The Microsporidia EF-1α data set is one of the prime examples showing that discrete rate shifts in some lineages can cause a long branch attraction bias in phylogenetic reconstruction (Inagaki et al. 2004). It is interesting to see whether the tests developed here can detect heterotachy for this EF-1α data set in the clade of Microsporidia versus the other eukaryotic and archaeal clades. Using either the true Microsporidia–Fungi clade tree or the incorrect Microsporidia–Archaea clade tree of EF-1α to simulate null distributions of the test statistics under RAS, the *w* test was not able to reject RAS ($P = 0.16$ and 0.21, respectively, based on the two competing trees), but the *w*′ test rejected RAS ($P = 0.05$ and 0.05, respectively) and the correlation test also rejected RAS ($P = 0.00$ and 0.00, respectively). This example shows that the *w*′ and correlation tests have more power in detecting heterotachy than the original *w* test for the Microsporidia EF data.

Based on a bivariate model, Susko et al. (2002) designed maximum likelihood–based orthogonal regression and parametric bootstrapping methods to analyze rate differences in the eEF-1α and aEF-1α data set and in the paralogous HBS1 and eRF3 data set. They found that there was a significant rate difference between the two subtrees of the eEF-1α and aEF-1α under both methods, whereas a marginally significant rate difference was detected in the second data set (HBS1 and eRF3) only with the parametric bootstrapping method. Here we applied the *w*, *w*′, and the correlation tests to the two data sets to distinguish them from null distributions of the three test statistics for data simulated under an RAS model. For simulating the null RAS distributions, we used the α parameter estimated from the total data sets or the average α from the two subtrees for each data set, respectively. Using either α for simulations, both *w* and *w*′ tests were not able to reject RAS for both data sets (see table 2 for *P* values). The

Table 2. The *w*, *w*′, and Correlation Tests on Four EF-Related Data Sets

Data Set	Number of Taxa in Clade 1	Number of Taxa in Clade 2	Number of Sites	*P* Value of Null Hypothesis (RAS)		
				w	*w*′	Pearson
EF-Tu/EF-1α	13 bacteria	17 eukaryotes	380	0.002[a]	0.00[a]	0.00[a]
EF-1α	3 Microsporidia	19 eukaryotes and 2 archaea	349	0.16	0.05[a]	0.00[a]
EF-1α	13 archaea	27 eukaryotes	269	0.73	0.49	0.00[a]
HBS1/eRF3	13 HBS1	17 eRF3	269	1.00	1.00	0.42

[a] The test rejected RAS.

correlation test rejected RAS for the third data set (eEF-1α and aEF-1α) but not the fourth one (HBS1 and eRF3).

Discussion

The foregoing results on both simulated and real data suggest that the *w* test sometimes lacks the sensitivity required to detect heterotachy when there are large numbers of taxa in two monophyletic groups. It is intriguing to note that Ané et al. (2005) found that the power of the *w*-based covarion tests increases with the number of the taxa per clade. However, their simulations only considered small data sets (2–16 taxa per clade). Our simulations show that it is for the biggest data sets (20 taxa per clade) that the *w* test has the least power. Furthermore, the power is especially low when the edge lengths within the subtrees are large. This is expected as in these cases both groups will likely vary at most sites. Therefore, few sites will show the N_3 and N_4 patterns leaving the *w* statistic powerless. The w' test, however, incorporates the entropy difference at sites so that covarion sites which vary in both groups can be detected and is thus still able to detect non-RAS processes in data of large numbers of taxa. The correlation method makes full use of the variability at sites between two subtrees and has even more power to detect non-RAS processes in large and very large data sets (greater than 300 taxa) than the w' test.

The *w* test was originally designed to distinguish the RAS process and the classic covarion process envisioned by Fitch and Markowitz (1970) and modeled by a Markov modulated rate-switching process (Tuffley and Steel 1998). In this study, we show that this test and its improved form (the w' test) also can detect other forms of heterotachy modeled by the general covarion model (Wang et al. 2007) and bivariate rate shift on splits (Susko et al. 2002). Under the bivariate rate shift model considered here, rates of the two subtrees are independently assigned, and thus less correlation between the rates at sites of the two subtrees is expected than from RAS-simulated data. This causes the correlation test to have very good power in distinguishing between bivariate data and RAS data. For the *w* test, increasing the α shape parameter used in simulating bivariate data will lead to low power. This effect, however, can be improved by using an α shape parameter derived by averaging α estimates from the two subtrees of the bivariate data to simulate the null RAS distribution for the *w* statistic (data not shown). The *w*, w', and correlation tests, however, cannot distinguish data simulated under one kind of the mixed branch length models (fig. 3) from RAS-generated data, and this cannot be remedied by using the averaged α approach. This is an example of a situation in which rates are varying across both lineages and sites, but the overall variation within groups does not show change across sites except for that due to the RAS distribution. It is exactly the type of setting where tests of this type will perform poorly. However, by subselecting taxa in the branch length mixture model (fig. 4), the power of the three tests and especially the w' and correlation tests increases dramatically, suggesting that, with some data exploration through taxa subselection, these tests can be used to detect heterotachy. Exactly how taxa should be subselected in the absence of prior information regarding partition-specific branch lengths remains a question for future research.

Unlike the branch length mixture model, both the covarion and the bivariate rate models are time-reversible stationary models of heterotachy that are expected to maintain a constant proportion of variable sites in all evolutionary lineages. However, this assumption may be overly simplistic when considering real data as proportions of variable sites (p_{var}) have been shown to vary in different lineages (Lockhart et al. 2006). This "changing p_{var} effect" has garnered a lot of attention with respect to the accuracy of phylogenetic inference methods in recent years (Gruenheit et al. 2008; Shavit Grievink et al. 2008, 2010). The simulation results suggest that the *w*, w', and correlation tests can effectively detect heterotachy in data that has lineage-specific changes in proportion of variable sites. Indeed, the lineage-specific p_{var} change in nonsister lineages in fact corresponds to a special case of the mixed branch length model with taxa subselection which, as discussed above, is shown to be detectable with the statistical tests introduced here. Although it is beyond the scope of this study, it will be of interest to investigate how these tests perform on different parameter settings for the changed p_{var} model and on more complex types of heterotachy such as having a combination of a rate-switching covarion process as well as changes in proportions of variable sites in the data.

One criticism of the w' and site entropy correlation tests as well as the original *w* test is that they do not explicitly account for the tree topology and edge lengths, which is indeed a weakness of the methods. However, by focusing on several potential clades of interest, these tests can provide a very fast way to detect heterotachy for any size of data, and the parametric bootstrapping method for obtaining the null distribution appears to in part control for topology and edge length effects. In terms of speed, covTests can analyze the simulated data sets of 300 taxa and 1,000 sites in 2.5 min on average on a computer with 3 GHz Intel Xeon processor and 16 GB RAM, which includes the bootstrapping approach. This is considerably faster than likelihood-based approaches, such as the PROCOV, Checkcov (Pupko and Galtier 2002), or Bivar (Susko et al. 2002). Furthermore, it appears that there is a good correlation between the site entropies and the site rates estimated with maximum likelihood. For example, we obtained the site rates estimated with Dist_Est (Susko et al. 2003) for the EF-1α/EF-Tu data and found them to be highly correlated with the site entropies ($R = 0.85$). The correlation between the site rates and entropies for the EF-Tu subset and for the EF-1α subset was even higher ($R = 0.94$ and 0.93, respectively). This high correlation may explain why the site entropy correlation test has better power than the *w* or w' tests in detecting heterotachy in the simulations and the EF data. Finally, the counting methods used in the calculation of *w*, w', and site entropies make no

distinction between the amino acid types, for instance, the variability of a site containing valine, leucine, and isoleucine will be treated same as valine, cysteine, and lysine. One way to account for differences in amino acid types is to recode the amino acids according to physicochemical properties. We have applied a recoding method (Susko and Roger 2007), based on Dayhoff physicochemical groups, into the w, w', and Pearson correlation analyses of several EF data sets. This indeed improved the power of the w tests but resulted in little change or slightly worse power for the w' and correlation tests.

Acknowledgments

We thank Tal Pupko, another reviewer, and Associate Editor Barbara Holland for critical comments on the manuscript. This work was supported by Discovery grants awarded to A.J.R. and E.S. by the Natural Sciences and Engineering Research Council of Canada. A.J.R. acknowledges support from the Canada Research Chairs Program. H.-C.W. is currently supported by a Centre for Comparative Genomics and Evolutionary Bioinformatics postdoctoral fellowship from the Tula Foundation.

References

Ané C, Burleigh J, McMahon M, Sanderson M. 2005. Covarion structure in plastid genome evolution: a new statistical test. *Mol Biol Evol*. 22:914–924.

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.

Dorman KS. 2007. Identifying dramatic selection shifts in phylogenetic trees. *BMC Evol Biol*. 7(Suppl 1):S10.

Fitch W, 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*. 4:479–593.

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

Gaucher EA, Miyamoto MM, Benner SA. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc Natl Acad Sci U S A*. 98:548–552.

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.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol*. 16:1664–1674.

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A*. 101:12957–12962.

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

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

Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A*. 98:14512–14517.

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

Kolaczkowski B, Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol Biol Evol*. 25:1054–1066.

Lockhart PJ, Huson D, Maier U, Fraunholz M, Van De Peer Y, Barbrook A, Howe C, Steel M. 2000. How molecules evolve in eubacteria. *Mol Biol Evol*. 17:835–838.

Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc Natl Acad Sci U S A*. 93:1930–1934.

Lockhart PJ, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. *Mol Biol Evol*. 23:40–45.

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.

Lopez P, Forterre P, Philippe H. 1999. The root of the tree of life in the light of the covarion model. *J Mol Evol*. 49:496–508.

Pagel M, Meade A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos Trans R Soc Lond B Biol Sci*. 363:3955–3964.

Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput Biol*. 4:e1000214.

Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc Biol Sci*. 269:1313–1316.

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

Shavit Grievink LS, Penny D, Hendy MD, Holland BR. 2008. LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. *BMC Evol Biol*. 8:317.

Shavit Grievink LS, Penny D, Hendy MD, Holland BR. 2010. Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. *Syst Biol*. 59:288–297.

Siltberg J, Liberles DA. 2002. A simple covarion-based approach to analyse nucleotide substitution rates. *J Evol Biol*. 15:588–594.

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

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

Strope CL, Abel K, Scott SD, Moriyama EN. 2009. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen Version 2.0. *Mol Biol Evol*. 26:2581–2593.

Studer RA, Robinson-Rechavi M. 2010. Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution. *Mol Biol Evol*. 27:2618–2627.

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, Inagaki Y, Field C, Holder ME, Roger AJ. 2002. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Mol Biol Evol*. 19:1514–1523.

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

Tuffley C, Steel MA. 1998. Modelling the covarion hypothesis of nucleotide substitution. *Math Biosci*. 147:63–91.

Wang HC, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol Biol.* 8:331.

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

Wang HC, Susko E, Roger AJ. 2009. PROCOV: maximum likelihood estimation of protein phylogeny under covarion models and site-specific covarion pattern analysis. *BMC Evol Biol.* 9:225.

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. 2009. General heterotachy and distance method adjustments. *Mol Biol Evol.* 26:2689–2697.

Zhou Y, Rodrigue N, Lartillot N, Philippe H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evol Biol.* 7:206.