# Difficulties in Testing for Covarion-Like Properties of Sequences under the Confounding Influence of Changing Proportions of Variable Sites

*Nicole Gruenheit,* Peter J. Lockhart,† Mike Steel,‡ and William Martin**

*Institute of Botany III, University of Düsseldorf, Düsseldorf, Germany; †Institute for Molecular BioSciences, Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and ‡Biomathematics Research Centre, Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand

The covarion (COV)-like properties of sequences are poorly described and their impact on phylogenetic analyses poorly understood. We demonstrate using simulations that, under an evolutionary model where the proportion of variable sites changes in nonadjacent lineages, log likelihood values for rates across site (RAS) and COV models become similar, making models difficult to distinguish. Further, although COV and RAS models provide a great improvement in likelihood scores over a homogeneous model with these simulated data, reconstruction accuracy of tree building is low, suggesting caution when it is suspected that proportions of variable sites differ in different evolutionary lineages. We study the performance of a recently developed contingency test that detects the presence of COV-type evolution modified for protein data. We report that if proportions of variable sites ($p_{var}$) change in a lineage-specific manner such that their distributions in different lineages become sufficiently nonoverlapping, then the contingency test can incorrectly suggest a homogeneous model. Also of concern is the possibility of different proportions of variable sites between the groups being studied. In a study of chloroplast proteins, interpretation of the test is found to be susceptible to different partitioning of taxon groups, making the test very subjective in its implementation. Extreme intergroup differences in the extent of divergence and difference in proportions of variable sites could be contributing to this effect.

## Introduction

Although sequence evolution is a temporally and spatially heterogeneous process, sequence evolution is typically described by a homogenous, stationary, time reversible model (Liò and Goldman 1998). Within this framework, improved phylogenetic estimates have often been obtained when site-specific properties of sequences have been modeled assuming that some sites are invariable (Adachi and Hasegawa 1995; Lockhart et al. 1996), nonindependent (von Haeseler and Schoniger 1998), and/or evolving with a discrete number of rate classes according to a gamma distribution (rates across site [RAS] models: Uzzel and Corbin 1971; Rzhetsky and Nei 1994; Yang 1994; Waddell et al. 1997).

More recently, a number of covarion (COV) (Fitch and Markowitz 1970) models have been implemented for phylogenetic analyses (Galtier 2001; Huelsenbeck 2002; Guindon et al. 2004; Wang et al. 2007), and these COV models have been found to provide further improvement over RAS models in terms of the relative fit to sequence data. This is presumably because these models capture a component of temporal heterogeneity in the evolutionary process—that is, unlike RAS models, they allow the substitution properties of a site to change over a time in a lineage-specific fashion. Under COV models, a site is free to switch back and forth between variable and invariable states along a branch.

In the COV model of Tuffley and Steel (1998), a site in a sequence may be either variable or invariable, and the state may differ in different lineages. All sites that are variable, evolve under the same substitution process (e.g., JC69, HKY85, etc.) and at the same rate. The COV model of Huelsenbeck (2002) extends the Tuffley and Steel model by allowing there to be a discrete number of rate classes for the variable state. Under this model, a site can switch between the OFF state and one of the variable rate classes but not between the different variable rate classes. A third COV model is that of Galtier (2001). In this model, there is a discrete number of rate classes for the variable state. A site can switch between these rate classes. Under this model, there is no OFF state. Most recently, Wang et al. (2007) have combined these 2 latter models and produced a general model (one in which there can be a switch between all variable states and an OFF state).

All these COV models are stationary time reversible models and have an expectation that the proportion of variable sites is the same in all evolutionary lineages. However, this assumption can be overly restrictive as proportions of variable sites, $p_{var}$, have been inferred to vary in lineage-specific ways (Lockhart et al. 1996, 2006; Lopez et al. 2002). This property of sequence evolution can lead to topological biases that will mislead tree building (Lockhart et al. 1996, 2006). With some proteins, changes in $p_{var}$ can be explained by lineage-specific differences in functional and structural constraints, due to differential loss/gain of functions ancillary to the core function of specific molecules (Susko et al. 2002; Inagaki et al. 2004; Guo and Stiller 2005).

Improving substitution models for phylogenetic analysis requires accurate tests to quantify the extent and nature of substitution model misspecification. A number of tests have been proposed to characterize COV-like substitution properties. However, as we illustrate using simulated and real data, interpretation from these tests need to be made cautiously, particularly when $p_{var}$ is not constant across the underlying phylogeny. Our findings highlight the need for improved analytical methods for studying the COV-like properties of sequences.

Key words: covarion, phylogenetics, chloroplast proteins, contingency test.

E-mail: nicole.gruenheit@uni-duesseldorf.de.

*Mol. Biol. Evol.* 25(7):1512–1520. 2008
doi:10.1093/molbev/msn098
Advance Access publication April 18, 2008



## Materials and Methods

### Maximum Likelihood Analyses

Within a maximum likelihood framework, log likelihood scores can be used to evaluate the relative fit of COV,

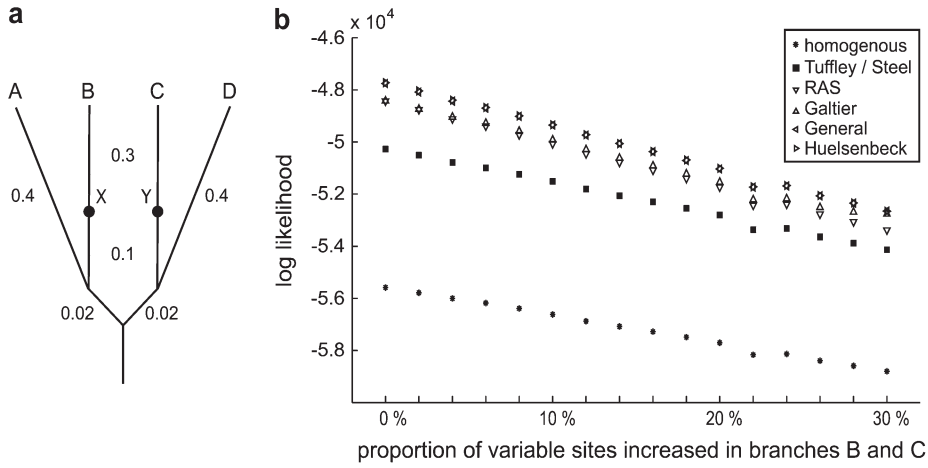


FIG. 1.—$P_{var}$ influences log likelihood. (*a*) Phylogenetic tree used for simulating alignments under a TS model. Relative branch lengths are indicated. A dot marks the point where the proportion of variable sites is increased in the outer branches leading to taxa B and C; at that point, 0–30% of the invariable sites are switched on. (*b*) Mean log likelihood for the simulated alignments according to the tree they were simulated on and 6 different models as indicated in the inset at upper right. The highest standard deviation (not plotted) for log likelihood among any 100 replicates is 329.43 found for the general model at $p_{var} = 0.2$.

RAS, and homogeneous models to sequence data. The best scores can then be used to identify the best substitution model for tree building. To examine the accuracy of this approach under conditions that might approximate the biological complexity expected with empirical data, we have examined the scores obtained when sequences are simulated under what we call a Tuffley and Steel (1998) + invariable site + switch (TS + I + S) model. In this model, a proportion of sites is specified as invariable and a proportion of sites is evolving under a TS model. At specified positions in the tree, a proportion of the specified invariable class switch to the TS class of sites, and some of the TS class of sites may also switch to the invariable class. On a 4-taxon tree that contains 2 switch positions (as in X and Y in fig 1), this model produces data that are identical to a phylogenetic mixture of 8 classes of TS model (each class has the same topology but different branch length), as described in table 1. This mixture representation is possible because a site that becomes invariable in various regions of the tree, but whose evolution is otherwise covered by the TS model, still follows a TS model but with branch lengths set to zero over the regions where the site is invariable.

However, in the special case of convergent increase in variable sites in nonadjacent lineages, where at the 2 switch positions 1) the only change in class is from invariable to TS sites and 2) the sites that change class are the same, the phylogenetic mixture reduces from 8 to just 3 classes of TS model. The mixture allowed us to specify the proportion of sites belonging to the invariable class and to specify the proportion of sites that switch from this class to the TS class in the nonadjacent lineages. Sequences 10,000 nt in length were simulated with Seq-Gen-Aminocov (Rambaut and Grassly 1997) specifying a Tuffley and Steel (1998) model wherein the variable states evolve according to a Jukes and Cantor (1969) rate matrix. Table 1 shows the calculation of relative partition sizes for mixtures that describe a 4-taxon tree (with relative branch lengths) wherein 1) the ancestral proportion of variable sites is 0.2; 2) the

**Table 1**

| Partitions | Frequency | Tree Mixtures for Simulating Tree with Changing $p_{var}$ Used in figure 1 |
|---|---|---|
| I | V(1 − qX − qY − qXY) | ([TaxonA:0.4,TaxonB:0.4]:0.02,[TaxonC:0.4,TaxonD:0.4]:0.02) |
| II | VqX | ([TaxonA:0.4,TaxonB:0.1]:0.02,[TaxonC:0.4,TaxonD:0.4]:0.02) |
| III | VqY | ([TaxonA:0.4,TaxonB:0.4]:0.02,[TaxonC:0.1,TaxonD:0.4]:0.02) |
| IV | VqXY | ((TaxonA:0.4,TaxonB:0.1):0.02,(TaxonC:0.1,TaxonD:0.4):0.02) |
| V | I(1 − pX − pY − pXY) | ([TaxonA:0,TaxonB:0]:0,[TaxonC:0,TaxonD:0]:0) |
| VI | IpX | ([TaxonA:0,TaxonB:0.3]:0,[TaxonC:0,TaxonD:0]:0) |
| VII | IpY | ([TaxonA:0,TaxonB:0]:0,[TaxonC:0.3,TaxonD:0]:0) |
| VIII | IpXY | ([TaxonA:0,TaxonB:0.3]:0.0,[TaxonC:0.3,TaxonD:0]:0.0) |

NOTE.—A rooted 4-taxon tree on which there is a change in the proportion of variable sites in 2 nonadjacent lineages (fig. 1) can be described in the general case by a mixture of 8 trees (same topology different branch lengths). For the simple case of convergent increase in $p_{var}$ in 2 nonadjacent lineages, only 3 trees need to be considered (I, V, and VIII). The positions X and Y at which there is an increase in $p_{var}$ (a switch of sites from the invariable to the TS class) are specified by the edge lengths. The ancestral $p_{var}$ and extent of change in $p_{var}$ in the nonadjacent lineages are specified by the partition sizes. These can be calculated for expected changes of $p_{var}$ as shown: where V and I = proportion of variable and invariable sites, respectively, at the root; pX = proportion of invariable sites that become variable at X but remain invariable at Y; pY = proportion of invariable sites that become variable at Y but remain invariable at X; pXY = proportion of invariable sites that become variable at both X and Y; qX = proportion of variable sites that become invariable at X but remain invariable at Y; qY = proportion of variable sites that become invariable at Y but remain variable at X; and qXY = proportion of variable sites that become invariable at both X and Y.

terminal branches have a length 0.4; and 3) where at a distance of 0.1 (at points *x* and *y*) along the branches to taxa B and C, the proportion of sites undergoing a Tuffley and Steel process is increased. $P_{var}$ was increased in increments of 2% up to 30% of the invariable class so that the overall $p_{var}$ in B and C ranged from 20% (no increase) to 50% (30% increase). For sites in the Tuffley and Steel class, a switching rate setting of 0.1 was used.

In our study for each of the increments, 100 replicates were generated, and each simulated alignment was analyzed using Procov1.3 (Wang et al. 2007). The following models were compared using the standard optimization files without reestimation of the branch lengths: homogeneous, Tuffley and Steel (1998), RAS (Yang 1994), Galtier (2001), General (Wang et al. 2007), and Huelsenbeck (2002). For each alignment and model, the log likelihood was extracted using a Perl script. The mean for each parameter was calculated and plotted using matlab, the standard deviations for each set of 100 replicates were very narrow (<0.1% of the mean in all cases) and hence were not plotted.

Trees were reconstructed for the simulated data sets using Paup* (Swofford 2003; maximum likelihood: lset nst = 1 basefreq = equal; lset tratio = 0.5 pinv = 0 rates = gamma shape = estimate; hsearch start = stepwise swap = tbr status = no nbest = 1; parsimony: hsearch start = stepwise swap = tbr status = no nbest = 1; parsimony: hsearch start = stepwise swap = tbr status = no nbest = 1) and MrBayes (Ronquist and Huelsenbeck 2003; lset nst = 1 covarion = yes; mcmc nruns = 1 ngen = 250000 samplefreq = 100 filename = run1.nex; sumt burnin = 400).

## Contingency Tests

Another approach to test whether a collection of sites in a multiple sequence alignment exhibit COV-type evolutionary properties is the contingency test developed by Lockhart et al. (1998), which is based on the test statistic *W*. This compares substitution differences between 2 groups of sequences

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

where $N_5$ is the number of sites that have varied in both groups, $N_3$ and $N_4$ are the numbers of those sites that have varied in one group but not in the other, and $N$ is the total number of sites. Site patterns referred to as $N_1$ and $N_2$ (Lockhart et al. 1998) are not relevant here. $N_1$ site patterns have the same residue in both groups. $N_2$ sites are polymorphic between but not within groups. *W* compares the fraction of varied sites in each group and the extent to which these sites overlap with sites that have varied in both groups.

$N_3$ or $N_4$ (syn. type 3 or type 4) sites should be less frequent if sequences are evolving according to a RAS model than if the sites are evolving in a manner that approximates a COV model (Lockhart et al. 1998). If in real data there are more $N_3$ and $N_4$ sites than expected to occur by chance under a RAS model, this would constitute evidence for deviation from the assumptions of a RAS model and possibly evidence for a COV modus of sequence evolution (Lockhart et al. 1998).

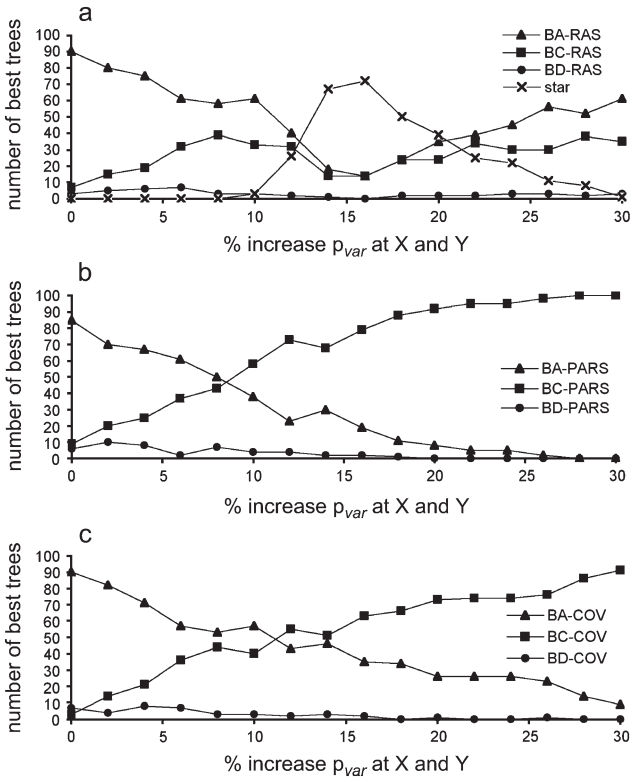


FIG. 2.—Phylogenetic reconstruction accuracy for increased proportions of variable sites. Data were simulated on the tree shown in figure 1 with incremental increases in $p_{var}$ as indicated on the abcissa, and the phylogeny was inferred with 3 different methods. (*a*) Maximum likelihood inference (RAS). (*b*) Parsimony inference (PARS). (*c*) Bayesian inference (COV). Reconstruction accuracy for all 3 methods drops markedly with an increase of only ~12% of invariable sites switching to the Tuffley and Steel site class at points *x* and *y* shown in figure 1. Only the maximum likelihood inference method delivered unresolved (star) trees. BA, BC, and BD designate the 3 possible topologies, respectively.

Ané et al. (2005) improved upon this test by providing a more rigorous means for obtaining expectations for the test statistic *W* under 3 different models of evolution 1) a homogeneous model, wherein different sequence positions are equally variable; 2) a RAS model, wherein some sites are evolving faster than other sites; and 3) a Tuffley and Steel (1998) COV model. In doing this, they noted that *W* predicts that sites that are varied in one group are likely to be varied in other groups under RAS and COV models but not under a homogeneous model. A RAS model predicts a strong degree of correlation and a COV model a weaker degree of correlation. Under a homogenous model, the *W* statistic is statistically zero. It is positive under a COV model and even more positive under a RAS model. The Ané et al. test uses simulation to interpret values of *W* in terms of support for each of the 3 models.

It does this by first examining whether there is evidence to reject a homogeneous model of sequence evolution in favor of a heterogeneous model. If so, it then examines whether there is evidence to reject a RAS model in favor of a more complex model of substitution. That is, if the derived *W* differs significantly from the expected

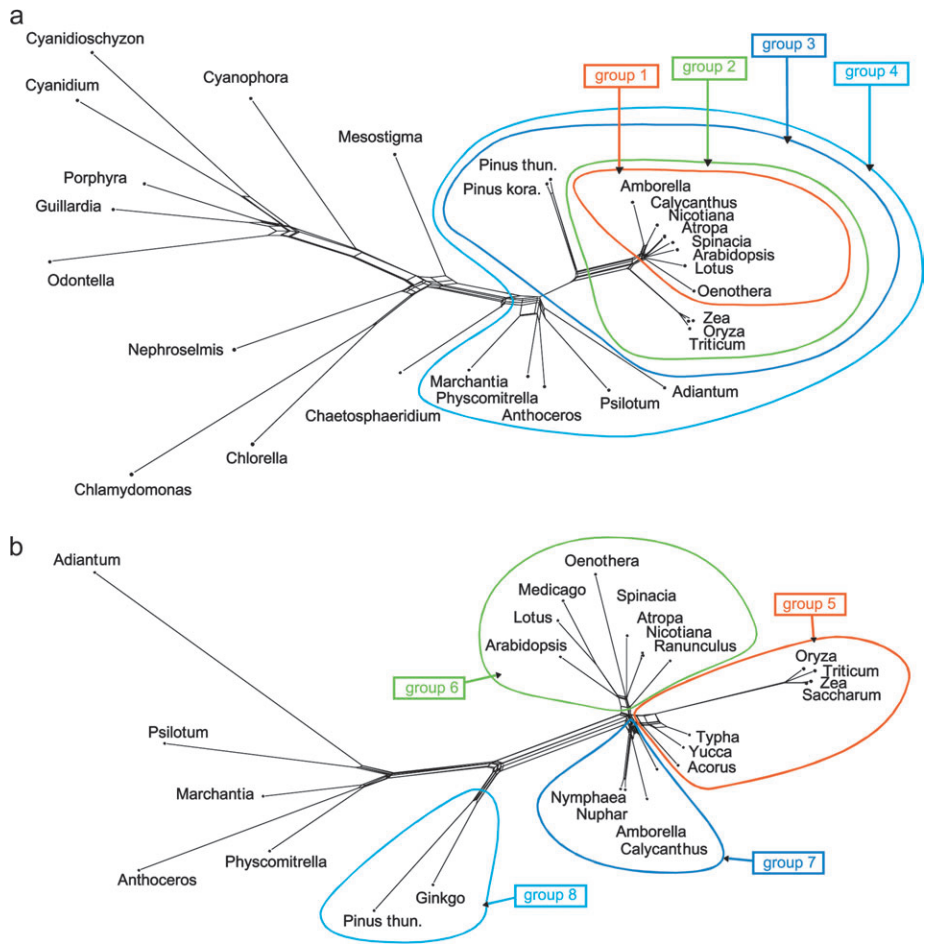


FIG. 3.—Comparisons used in the present study. (*a*) Neighbor-Net (Bryant and Moulton 2004) of 45 concatenated alignments of chloroplast proteins representing all taxa of the old data set. Marked with colored boxes are the groups used in the analyzed comparisons. Groups 1 (red, eudicots) and 2 (green, angiosperms) were proposed in Ané et al. (2005). In addition to those, 2 more groups were chosen. Group 3 (blue) contains all angiosperms and 2 gymnosperms, group 4 (turquoise) contains group 3 and all mosses and ferns. (*b*) Neighbor-Net of 58 concatenated alignments of chloroplast proteins representing taxa in the second data set. Groups used in the comparisons are indicated.

distribution of the $W$ under a RAS model, the nucleotide or protein sequence is inferred to have evolved under a RAS + COV model. Ané et al. (2005) used this test to infer that a large proportion of proteins encoded in chloroplast genomes evolve according to a RAS + COV model.

We have implemented the method of Ané et al. (2005) for analyzing protein sequences and used it to reexamine chloroplast genome sequences also studied by Ané et al. The sequences used are from *Acorus calamus* (NC_007407), *Adiantum capillus-veneris* (NC_004766), *Amborella trichopoda* (NC_005086), *Anthoceros formosae* (NC_004543), *Arabidopsis thaliana* (NC_000932), *Atropa belladonna* (NC_004561), *Calycanthus floridus* (NC_004993), *Chaetosphaeridium globosum* (NC_004115), *Chlamydomonas reinhardtii* (NC_005353), *Chlorella vulgaris* (NC_001865), *Cyanidioschyzon merolae* (NC_004799), *Cyanophora paradoxa* (NC_001675), *Epifagus virginiana* (NC_001568), *Ginkgo biloba* (DQ069337–DQ069702), *Guillardia theta* (NC_000926), *Lotus corniculatus* (NC_002694), *Marchantia polymorpha* (NC_001319), *Medicago truncatula* (NC_003119), *Mesostigma viride* (NC_002186), *Nephroselmis olivacea* (NC_000927), *Nicotiana tabacum* (NC_001879), *Nuphar advena* (DQ069337–DQ069702), *Nymphaea alba* (NC_006050), *Odontella sinensis* (NC_001713), *Oenothera elata* (NC_002693), *Oryza sativa* (NC_001320), *Physcomitrella patens* (NC_005087), *Pinus koraiensis* (NC_004677), *Pinus thunbergii* (NC_001631), *Porphyra purpurea* (NC_000925), *Psilotum nudum* (NC_003386), *Ranunculus macranthus* (DQ069337–DQ069702), *Saccharum officinarum* (NC_006084), *Spinacia oleracea* (NC_002202), *Triticum aestivum* (NC_002762), *Typha latifolia* (DQ069337–DQ069702), *Yucca schidigera* (DQ069337–DQ069702), and *Zea mays* (NC_001666).

Sequences were aligned using ClustalW (Thompson et al. 1994), and all gapped sites were removed. To obtain a phylogenetic overview of the data set, sequences were concatenated, LogDet distances were computed with LDDist (Lake 1994; Lockhart et al. 1994; Thollesson 2004) from which phylogenetic networks were constructed with Neighbor-Net as implemented in splitstree 4 (Huson and Bryant 2006). A Java program was written to count the different types of sites and is available upon request. For each alignment, the user gets the numbers of type 1, 2, 3, 4, and 5 sites. Sites with gaps have been ignored.

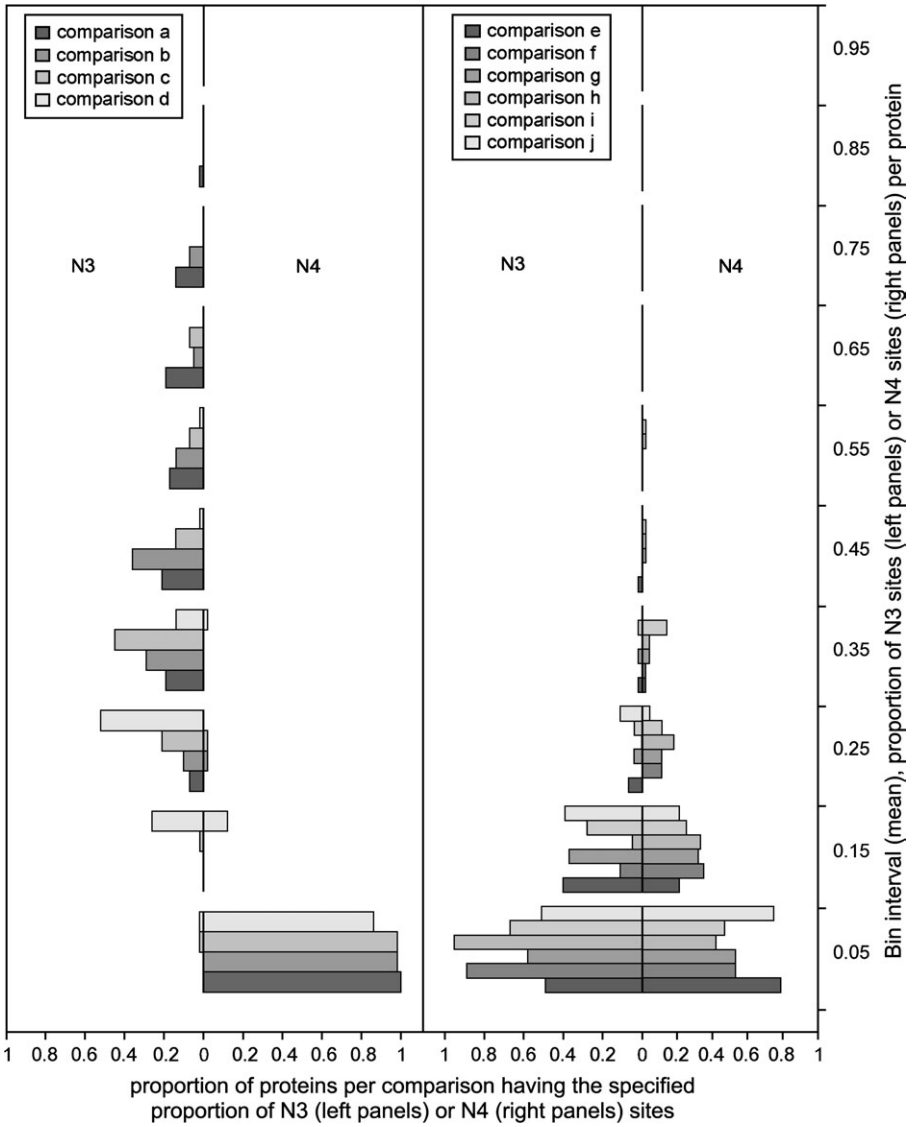


FIG. 4.—Influence of groups compared upon relative proportions of inferred $N_3$ and $N_4$ sites in chloroplast-encoded proteins. Proportion of proteins in different group comparisons is given on the *y* axis. Proportions of $N_3$ and $N_4$ sites per protein are indicated on the *x* axis. Groups compared are shown in figure 3, comparisons (inset) are listed in table 1. (*a*) Ingroup–outgroup comparisons. (*b*) Comparisons of monophyletic groups.

## Results

### Maximum Likelihood Analyses

We have investigated the extent to which time reversible substitution models describe the evolution of sequences that have evolved under a Tuffley and Steel (1998) model where the proportion of variable sites, $p_{var}$ does not remain constant across all lineages. Figure 1 shows the relative fit (log likelihood scores) of homogeneous, RAS, and 3 COV models to these simulated data. When the sequences have evolved under the comparatively simple Tuffley–Steel model, more complicated COV models nevertheless gave improved likelihood scores both when $p_{var}$ was constant across all lineages and when $p_{var}$ was incrementally increased to 0.3 in 2 nonadjacent lineages. Changes in $p_{var} > 0.16$ resulted in differences of log likelihood for replicates for a given heterogeneous model that exceeded the differences among different heterogeneous models for a given $p_{var}$. As further shown in figure 1, log likelihood values for the different substitution models began to converge as $p_{var}$ increased in nonadjacent lineages. For these data, tree building exhibited low reconstruction accuracy (fig. 2). A change in merely 8–12% of the invariable sites becoming variable in nonadjacent lineages caused sufficient topological distortion (long branches leading to sequences B and C) to mislead maximum likelihood (lset = 1, assumed discrete gamma, estimated alpha; fig. 2*a*), parsimony (fig. 2*b*), and Bayesian (lset = 1, assumed Huelsenbeck COV model, estimated switching rates; fig. 2*c*) tree building. Thus, although RAS and heterogeneous COV models provided an improved fit to the sequences, as $p_{var}$ increased in nonadjacent lineages it became more difficult to distinguish among the different heterogeneous models, and neither RAS nor a COV model was sufficient to allow for reliable

**Table 2**
**Comparisons of Group A versus Group B and Proportions of $N_3$ and $N_4$ Sites among All Sites**

| Comparison | Group A | Group B | p$N_3$ | p$N_4$ | p$N_3$/p$N_4$ |
|---|---|---|---|---|---|
| a | 1 | All others | 0.498 | 0.005 | 99 |
| b | 2 | All others | 0.408 | 0.019 | 22 |
| c | 3 | All others | 0.345 | 0.028 | 12 |
| d | 4 | All others | 0.233 | 0.049 | 4.7 |
| e | 5 | 6 | 0.108 | 0.067 | 1.6 |
| f | 5 | 7 | 0.045 | 0.109 | 0.41 |
| g | 5 | 8 | 0.077 | 0.112 | 0.68 |
| h | 6 | 7 | 0.035 | 0.141 | 0.25 |
| i | 6 | 8 | 0.067 | 0.144 | 0.46 |
| j | 7 | 8 | 0.091 | 0.063 | 1.4 |

NOTE.—Group designations refer to those shown in figure 3*a* and *b*. The columns p$N_3$ and p$N_4$ indicate the total number of $N_3$ and $N_4$ sites, respectively, divided by the number of all sites compared. In comparisons a–d, 9,490 site patterns in 42 alignments were investigated (data set #1), in comparisons e–j, 11,584 site patterns in 57 alignments were investigated (data set #2).

reconstruction of the true phylogeny with only moderate increases in $p_{var}$. This underscores a significant and seldom examined effect of $p_{var}$ in phylogenetic inference.

## Contingency Tests

Characterization of the COV-like properties of sequence data can also be made using contingency tests. The test of Ané et al. overcomes problems of interpreting the *W* statistic with real data that were not solved by Lockhart et al. (1998). In doing so, these authors also described the impact that taxon sampling is expected to have on the power of the test. They studied this for the case of 2 monophyletic groups in terms of the edge lengths within (*t*) and between (*T*) compared clades. However, in implementing the test, these authors compared a monophyletic group and a paraphyletic group. Although, the test is still validly applied in this case, we report that the expectations for performance of the test differ from that described when 2 monophyletic groups are compared. In demonstrating this, 2 different data sets were analyzed. The first data set (#1) comprised 29 land plants and algae (fig. 3*a*) and 42 chloroplast proteins. In the second data set (#2), there were 26 land plants (fig. 3*b*) and 57 chloroplast proteins. The contingency test of Ané et al. (2005) was adapted to investigate protein instead of nucleotide sequences (source code available upon request). As in Ané et al. (2005), we identified groups for comparison: (group 1) eudicotyledons, (group 2) angiosperms, and (group 3) angiosperms and gymnosperms. We also considered (group 4) angiosperms, gymnosperms, moss, and ferns. In the implementation of the test of Ané et al. (2005), each of the first 3 groups was compared against the rest of the data set. In our implementation, 2 monophyletic groups were compared. The $N_3$ and $N_4$ sites in the alignment were counted for each comparison using a Java script and plotted as histograms (fig. 4).The proportions of $N_3$ and $N_4$ sites among all sites are shown in table 2.

A striking feature of the aligned sequence data are the different proportions of $N_3$ and $N_4$ sites among different groups of sequences. In comparisons of a monophyletic versus paraphyletic group, the number of $N_3$ sites greatly exceeds the number of $N_4$ sites. All proteins had at least 20% $N_3$ sites, and in 16% of the proteins >70% of all sites were $N_3$, whereas no protein had an $N_4$ site (fig. 4*a*). In some proteins, more than 80% of all sites were $N_3$ or $N_4$ sites. In the comparison of 2 monophyletic groups, far fewer $N_3$ and $N_4$ sites were found and a considerable greater balance between the numbers of $N_3$ and $N_4$ sites was observed (fig. 4*b*). Most proteins had <5% of either $N_3$ or $N_4$ sites, the maximum number of $N_3$ or $N_4$ sites lies between 40% and 50%. Ané et al. (2005) detected 21 proteins that were inferred to reject the RAS model using nucleotide site patterns. Using the same groups and amino acid site patterns (instead of nucleotide site patterns), we found that 28/42 (67%) of the proteins tested (data set #1) would reject the RAS model in all comparisons of the ingroup versus outgroup type. By contrast, only one protein out of 57 investigated (data set #2), rbcL, rejected RAS in all comparisons of monophyletic groups (table 3). Thus, balanced verses unbalanced sampling of sequences gave very different results in terms of evidence for COV-like properties of the sequences.

A further property of the test statistic *W* also suggests caution in its application. This is that *W* can become negative (or close to 0) when distributions of variable sites in the groups being compared become sufficiently different, as might happen if the spatial pattern of substitution differs from that expected under time reversible COV models. Thus, unexpected but nevertheless COV-like patterns could lead the *W* statistic to underestimate the heterogeneity of the substitution process. The expected value *w* of *W* can be written as:

$$w = p_{12} - p_1 p_2,$$

where $p_i$ is the probability that a site has varied in group 1 or 2 and where $p_{12}$ is the probability that a site has varied in group 1 and group 2. Under both the Tuffley–Steel model and the RAS model $w \geq 0$. However, if the distribution of variable sites has evolved in a more complex manner than envisaged by Tuffley–Steel, then it can be shown that $w \leq 0$. For example, consider a model where sites fall into 4 classes depending on whether they are variable or invariable in the 2 groups $G_1$, $G_2$, and let

$v_i$ = Proportion of variable sites in group $G_i$,
$v_{12}$ = Proportion of variable sites in groups $G_1$ and $G_2$,
$\pi_i$ = Probability that a site that is variable in $G_i$ is varied in $G_i$, and
$\pi_{12}$ = Probability that a site that is variable in $G_1$ and $G_2$ is varied in $G_1$ and $G_2$.

Then $p_i \approx \pi_i v_i$ and $p_{12} \approx \pi_{12} v_{12}$, and for a substitution process that is group based (e.g., Jukes and Cantor; Kimura 2P and Kimura 3ST models), we also have $\pi_{12} = \pi_1 \pi_2$ and $w \approx \pi_1 \pi_2 (v_{12} - v_1 v_2)$. If the proportion of variable sites increases in $G_2$ whereby the variable sites in $G_1$ are a subset of the variable sites in $G_2$, then the proportion of sites variable in both $G_1$ and $G_2$ will equal the proportion of sites variable in $G_1$, thus $v_{12} = v_1$ and $w \geq 0$ because $w \approx \pi_1 \pi_2 (v_{12} - v_1 v_2) = \pi_1 \pi_2 v_1 (1 - v_2) \geq 0$. However, if there is little, or in the extreme case, no overlap in the sites that are variable in $G_1$ and $G_2$, then *w* can take a negative value.

**Table 3**
**Proteins Rejecting a RAS Model at $P = 0.95$**

| Protein | Comparison | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | a | b | c | d | e | f | g | h | i | j |
| atpA | * | * | * | * | * | * | * | . | . | . |
| atpB | * | * | * | * | * | * | * | * | . | . |
| atpE | * | * | * | * | . | . | . | * | . | . |
| atpF | * | * | * | * | . | . | . | . | . | . |
| atpH | . | * | * | * | . | * | . | * | . | * |
| atpI | | | | | * | . | . | * | . | * |
| cemA | | | | | . | . | . | . | . | . |
| clpP | | | | | . | . | . | . | . | * |
| petA | * | * | * | * | . | * | . | * | . | . |
| petB | * | . | * | * | . | * | . | * | . | . |
| petD | * | * | * | * | . | . | . | . | . | . |
| petG | | | | | * | * | * | . | . | . |
| petL | | | | | . | . | . | . | . | . |
| petN | | | | | . | * | * | * | . | * |
| psaA | * | * | * | * | * | * | . | * | * | * |
| psaB | * | * | * | * | * | * | * | . | * | * |
| psaC | . | . | * | * | . | . | * | . | * | * |
| psaI | | | | | . | . | . | . | . | . |
| psaJ | * | . | * | * | . | . | . | . | . | . |
| psbA | . | * | * | * | . | . | . | . | * | . |
| psbB | * | * | * | * | * | * | . | * | . | . |
| psbC | * | * | * | * | . | . | . | * | * | * |
| psbD | * | * | * | * | * | . | . | . | . | . |
| psbE | . | . | * | . | . | . | . | . | . | . |
| psbF | * | . | . | . | . | . | . | . | . | . |
| psbH | * | * | * | * | . | . | . | . | . | . |
| psbI | | | | | . | . | . | . | . | . |
| psbJ | * | * | * | * | . | . | . | . | . | . |
| psbK | * | * | * | * | . | . | . | . | . | . |
| psbL | * | . | * | . | . | . | . | . | . | . |
| psbM | | | | | . | . | . | . | . | . |
| psbN | * | * | . | . | . | . | . | . | . | . |
| psbT | . | . | . | . | * | * | * | * | . | * |
| psbZ | | | | | . | . | . | . | . | . |
| rbcL | | | | | * | * | * | * | * | * |
| rpl14 | . | . | . | * | . | . | . | . | . | . |
| rpl16 | * | * | * | * | . | * | . | . | . | . |
| rpl2 | * | * | * | * | * | * | . | * | . | . |
| rpl20 | * | * | * | . | * | . | . | . | . | . |
| rpl32 | | | | | . | . | . | . | . | . |
| rpl33 | | | | | . | . | . | . | . | . |
| rpl36 | | | | | . | . | . | . | . | . |
| rpoB | * | * | * | * | . | . | . | * | . | . |
| rpoC1 | * | * | * | * | . | . | . | * | . | . |
| rpoC2 | * | * | * | * | * | . | . | * | . | . |
| rps11 | * | * | * | * | . | . | . | . | . | . |
| rps12 | * | . | . | . | . | . | * | . | . | . |
| rps14 | * | * | * | * | * | . | . | . | . | . |
| rps18 | . | . | . | . | . | . | . | . | . | . |
| rps19 | * | * | * | * | . | . | . | . | . | . |
| rps2 | * | * | * | * | . | . | . | . | . | . |
| rps3 | * | * | * | * | . | . | . | . | . | . |
| rps4 | * | * | * | * | * | . | . | . | . | . |
| rps7 | * | * | * | * | . | . | . | . | . | . |
| rps8 | * | * | * | * | . | . | . | . | . | . |
| ycf3 | * | * | * | * | * | . | . | . | . | . |
| ycf4 | | | | | . | . | . | . | . | * |

NOTE.—Comparisons a–j refer to those in table 2. An asterisk indicates that RAS was rejected at $P = 0.05$, a dot indicates that RAS was not rejected, and empty elements indicate that the comparison was not performed (gene not present or paralogous [red lineage rbcL] in at least one genome compared).

As a simple example, this could entail a hypothetical protein 100 amino acids in length. In $G_1$, the 30 N-terminal sites of this protein become variable but the 70 C-terminal sites remain constant, whereas in $G_2$, the 30 C-terminal sites of X become variable but the 70 N-terminal sites remain constant. In this case, $w \leq 0$ even though the proportion of variable sites in the 2 groups is similar or the same ($v_1 = v_2$) provided that $v_{12} < v_1^2$ (because if $v_1 = v_2$, then $w \approx \pi_1\pi_2(v_{12} - v_1v_2) = \pi_1\pi_2(v_{12} - v_1^2)$) because $w \approx \pi_1\pi_2(v_{12} - v_1v_2) = \pi_1\pi_2(0 - (0.3 \times 0.3)) < 0$.

## Discussion

For confidence in the reliability of tree building from highly diverged sequences, it is essential to develop low parameter substitution models that capture the heterogeneous complexity of sequence evolution. However, as we have illustrated, current methods need to be applied cautiously in characterizing the evolutionary properties of highly diverged sequences, and our current understanding of sequence evolution is limiting for model development. In this respect, it is important to note that tests of heterotachy, which we have not discussed (e.g., Lopez et al. 1999; Misof et al. 2002; Susko et al. 2002; Baele et al. 2006), while being informative are nevertheless not sufficient for developing models of sequence evolution. The reason is that different processes of change can lead to very similar patterns of heterotachy. These tests cannot distinguish an evolutionary model where there is a constant rate of evolution, but different proportions of variable sites in different lineages (the model studied by Lockhart and Steel 2005), from a model where there is the same proportion of variable sites in different lineages and lineage-specific rates of substitution (the scenario studied by Felsenstein 1978). This distinction is important because as demonstrated here when $p_{var}$ changes, model fitting can favor a model that does not improve phylogenetic accuracy. Contingency tests to identify COV-like properties may seem promising, but their implementation is problematic. Sampling of taxa can significantly impact on the outcome of the test and deciding upon an objective sampling criterion is not straightforward. In the present study, the contingency test of Ané et al. (2005) gave very different results depending on whether comparisons were made between 2 monophyletic groups or a monophyletic group and a paraphyletic group. Both comparisons are valid, but which result is correct? Further, it is unclear whether this difference is due to the much greater divergence among the paraphyletic species (this group containing algae, e.g., which have had much more time to evolve than sites in the eudicots; hence, many $N_3$ sites are expected even under a RAS model) or whether it is because the substitution properties in the algal sequences differ significantly from those in the higher plants (Lockhart et al. 2006; Rodriguez-Ezpelata et al. 2007).

A recent development in modeling substitution properties of sequences is to fit a mixture of substitution models to each site in an alignment of sequences (e.g., Pagel and Meade 2004; Lartillot et al. 2007). This approach can also be extended to fit a mixture of trees with different branch lengths to the sequences (e.g., Kolaczkowski and Thornton 2004; Zhou et al. 2007). There are issues of identifiability with complex mixture and COV models (Allman and Rhodes 2007), but potentially tests might be developed using such models to better characterize temporal heterogeneity in the evolution of sequences. Such developments will be important because although RAS models have generally improved phylogenetic inference, as we demonstrate here, they are unable to account for lineage-specific patterns of changing $p_{var.}$ They, and currently implemented COV models, are unable to account for the form of heterotachy that most likely describes the evolution of biological sequences, the further development of mixture models is of interest in this respect.

## Acknowledgments

We thank Simon Whelan, Andrew Roger, Ed Susko, John Rhodes, Liat Shavit, Simon Joly, Elizabeth Allman, Oliver Deusch, and Tal Dagan for helpful discussions and Microsoft (P.J.L.) and the Julius von Haast Fellowship Fund (W.M.) for research fellowships. This work was funded by the New Zealand Marsden Fund (P.J.L.) and the Deutsche Forschungsgemeinschaft (W.M.).

## Literature Cited

Adachi J, Hasegawa M. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J Mol Evol. 40: 622–628.

Allman ES, Rhodes J. 2007. The identifiability of tree topology for phylogenetic models. J Comput Biol. 13:1103–1113.

Ané C, Burleigh JG, MacMahon MM, Sanderson MJ. 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.

Bryant D, Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21:255–265.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading, Syst Zool. 27:401–410.

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

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

Guo Z, Stiller J. 2005. Comparative genomics and evolution of proteins associated with RNA polymerase II C terminal domain. Mol Biol Evol. 22:2166–2178.

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

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

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

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

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro H N, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

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

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

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7(Suppl 1):S4.

Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Res. 8:1233–1244.

Lockhart PJ, Larkum AWD, Steel MA, Waddell PJ, Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the

problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA. 93:1930–1934.
Lockhart PJ, Novis P, Milligan BG, Riden J, Rambaut A, Larkum AWD. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.
Lockhart PJ, Steel M. 2005. A tale of two processes. Syst Biol. 54:948–951.
Lockhart PJ, Steel M, 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.
Lockhart PJ, Steel M, Hendy M, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 11:605–612.
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.
Misof B, Anderson CL, Buckley TR, Erpenbeck D, Rickert A, Misof K. 2002. An empirical analysis of mt 16S rRNA covarion-like evolution of insects: site-specific rate variation is clustered and frequently detected. J Mol Evol. 55:460–469.
Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern heterogeneity in gene sequence or character-state data. Syst Biol. 53:571–581.
Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte-Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 13:235–238.
Rodriguez-Ezpelata N, Philippe H, Brinkmann H, Burkhard B, Melkonian M. 2007. Phylogenetic analyses of nuclear, mitochondrial and plastid multi-gene datasets support the placement of *Mesostigma* in the Streptophyta. Mol Biol Evol. 24:723–731.
Ronquist F, Huelsenbeck JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Rzhetsky A, Nei M. 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J Mol Evol. 38:295–299.
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.
Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4. Sunderland (MA): Sinauer.
Thollesson M. 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics. 20:416–418.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.
Tuffley C, Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci. 147:63–91.
Uzzel T, Corbin KW. 1971. Fitting discrete probability distributions to evolutionary events. Science. 172:1089–1096.
von Haeseler A, Schoniger M. 1998. Evolution of DNA or amino acid sequences with dependent sites. J Comput Biol. 5: 149–164.
Waddell PJ, Penny D, Moore T. 1997. Hadamard conjugations and modelling sequence evolution with unequal rates across sites. Mol Phylogenet Evol. 8:33–50.
Wang H-C, Spencer M, Susko E, Roger AJ. 2007. Testing for covarion-like evolution in protein sequences. Mol Biol Evol. 24:294–305.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39:306–314.
Zhou Y, Rodrigue N, Lartillot N, Philippe H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol Biol. 7:206.

Andrew Roger, Associate Editor

Accepted April 13, 2008