

Covariation Structure in Plastid Genome Evolution: A New Statistical Test

Cécile Ané,¹ J. Gordon Burleigh, Michelle M. McMahon, and Michael J. Sanderson

Section of Evolution and Ecology, University of California, Davis

Covariation models of molecular evolution allow the rate of evolution of a site to vary through time. There are few simple and effective tests for covariation evolution, and consequently, little is known about the presence of covariation processes in molecular evolution. We describe two new tests for covariation evolution and demonstrate with simulations that they perform well under a wide range of conditions. A survey of covariation evolution in sequenced plastid genomes found evidence of covariation drift in at least 26 out of 57 genes. Covariation evolution is most evident in first and second codon positions of the plastid genes, and there is no evidence of covariation evolution in third codon positions. Therefore, the significant covariation tests are likely due to changes in the selective constraints of amino acids. The frequency of covariation evolution within the plastid genome suggests that covariation processes of evolution were important in generating the observed patterns of sequence variation among plastid genomes.

Introduction

The ever-increasing amount of sequence data and the availability of new statistical and computational methods have motivated the development of progressively more complex models of evolution (Liò and Goldman 1998; Whelan et al. 2001; Holder and Lewis 2003). Though variation in the substitution rate across nucleotide or amino acid sites is commonly incorporated into evolutionary analyses (e.g., Yang 1994, 1996), variation in the rate of evolution of a site through time is rarely considered (Galtier 2001; Huelsenbeck 2002). This may be due to difficulties in implementing such a model and the lack of simple tests to determine if it would be appropriate. The covariation (for concomitantly variable codon) hypothesis of molecular evolution proposes that selective pressures on an amino acid or nucleotide site change throughout time, and therefore, a site's rate of evolution also changes (Fitch and Markowitz 1970; Fitch 1971). The term covariation sometimes specifically refers to protein sequences while covariotide refers to nucleotide sequences (Shoemaker and Fitch 1989); however, we will use the more general term covariation to refer to shifts in the substitution rate of any character. In the covariation hypothesis, the functional constraints of a site change through time, and a character that is functionally constrained within one lineage may not be constrained in another lineage. In an extreme case, a site may be either completely constrained or unconstrained and susceptible to substitutions. Under this model, there would be an excess of sites that are invariant in one part of a tree but variant in another. Different sites could be variant or invariant in different parts of the tree. Consequently, the first evidence of covariation-like patterns of evolution was based on detecting sites that had no variation among taxa in one clade and variation among taxa in another clade (e.g., Fitch and Markowitz 1970; Fitch 1971; Miyamoto and Fitch 1995; Lockhart et al. 1998).

Tuffley and Steel (1998; also see Penny et al. 2001) developed the first formal model of covariation evolution. In their model, the substitution process can be turned ON or OFF. Whenever a site is ON, it evolves according to some sub-

stitution process, and when a site is OFF, that site is invariant. The ON substitution process can be modeled with any reversible substitution rate matrix (see e.g., Swofford et al. 1996). The switches between ON and OFF are modeled as an additional stationary Markov process with two parameters, the ON equilibrium frequency σ and the average number of switches per substitution v . The transition matrix of the switch process has ON/OFF switching rate s_{01} and an OFF/ON switching rate s_{10} that are determined by σ and v . Specifically, the ON/OFF switching rate $s_{01} = \sigma v / (2(1 - \sigma))$, and the OFF/ON switching rate $s_{10} = v/2$. If $\sigma = 0$, all sites are always invariant, and if $\sigma = 1$, the sites always evolve according to the normal substitution model. Furthermore, if $v = 0$, there are no switches between ON and OFF or OFF and ON, and any given site will be either invariant or variable throughout the tree. As v converges on ∞ , the model of evolution resembles the normal substitution model. Such a model may incorporate variation in rates of evolution across sites, which is often modeled using a discrete gamma distribution (Yang 1994). In a model with rate variation across sites, sites may have different rates of evolution, but the rate of evolution for a single site remains constant throughout the tree. We will refer to a model with variable rates across sites as a RAS model, a covariation model as a COV model (COVArion), and a model with both among-site rate variation and a covariation evolution as a COV + RAS model. Huelsenbeck (2002) adapted the Tuffley and Steel (1998) covariation model to allow any reversible nucleotide model, including RAS models, to incorporate covariation evolution. Galtier (2001) developed a different covariation model that allows the rate at a site to vary over time and does not require that sites turn completely on or off.

There are few formal tests of covariation evolution in sequence data. Lockhart et al. (1998) developed a nonparametric test to detect covariation evolution under the Tuffley and Steel (1998) model and applied it to two loci, and another nonparametric covariation test detected evidence of covariation evolution in four protein-coding loci (Lockhart et al. 2000). Galtier (2001) used approximate likelihood estimates to perform likelihood ratio tests on two loci. Huelsenbeck (2002) developed an integrated likelihood ratio test for covariation evolution using MCMC and found evidence of covariation evolution in 9 out of 11 loci. Though likelihood ratio tests often perform well in evolutionary model testing (Posada and Crandall 2001), the likelihood ratio tests of covariation models are computationally complex and difficult to implement. The complexity of calculating the likelihood

¹ Present address: Department of Statistics, University of Wisconsin.

Key words: covariation, model testing, parametric bootstrap, plastid genome evolution.

E-mail: ane@stat.wisc.edu.

Mol. Biol. Evol. 22(4):914–924. 2005

doi:10.1093/molbev/msi076

Advance Access publication December 29, 2004

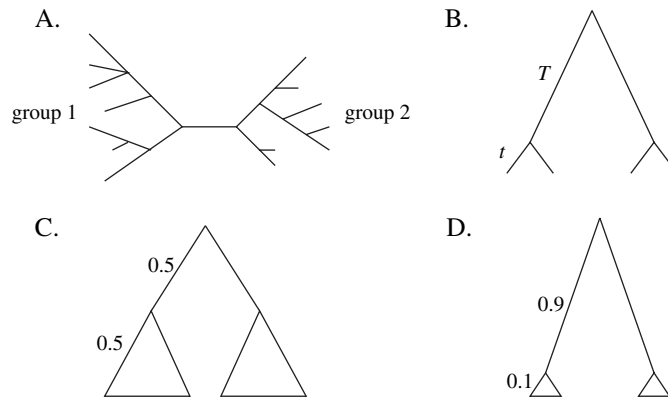


FIG. 1.—Example trees for heterogeneity and covariation tests. (A) is an unrooted tree with a bipartition separating two groups. (B) is an example of a four-taxon rooted tree. T is the distance from the root of each clade to the root of the tree, and t is the distance from each tip to the root of its clade. (C) and (D) represent trees used in the simulation experiments. The tree in (C) represents a tree with deep clades, and the tree in (D) represents shallow clades. The numbers on the branches are the average number of substitutions per site.

with a covariation model is due to bigger substitution matrices (8×8 instead of 4×4) that, moreover, need to be diagonalized more often during optimization. Additionally, the performance of covariation model testing methods is not well characterized. Therefore, it is not surprising that these tests are rarely utilized, and the importance of the covariation process in molecular evolution is largely unknown.

We describe two new tests for covariation evolution that are based on a test statistic from Tuffley and Steel (1998) and Lockhart et al. (1998). We examine the performance and power of the new tests under a wide range of conditions using simulations. Finally, we use the test to survey for the presence of covariation patterns of evolution in 57 genes obtained from complete plastid genomes.

Methods

Tests of the Covariation Hypothesis

The tests of covariation evolution are based on the Tuffley and Steel (1998) model of the covariation process, though the tests also account for rate variation among sites. The test statistics measure the independence of the substitution process between two groups of taxa. First, an unrooted topology is split into a bipartition. Each element of the bipartition is called a *group* (fig. 1A). The total number of sites in a sequence alignment is N . Let N_1 be the number of sites that vary in the first group, N_2 be the number of sites that vary in the second group, and N_{12} be the number of sites that vary in both groups. The probability p_1 that a site varies in the first group is estimated by N_1/N , and N_2/N is an estimate of the probability p_2 that a site varies in the second group. If site variation is independent in the two groups, then the probability p_{12} that a site varies in both groups would be the product $p_1 p_2$. The difference $w = p_{12} - p_1 p_2$ is a measure of the correlation of site variation in the two groups. Because N_{12}/N is an estimate of p_{12} , we construct the test statistic

$$W = N_{12}/N - N_1 N_2 / N^2.$$

If there is no rate variation across sites or within the tree and the underlying substitution process follows the

Kimura 3ST (Kimura 1981) or Jukes and Cantor (1969) model, then w is exactly 0 (Tuffley and Steel 1998) because in these models the probability that a site is variable within a group is not dependent on the ancestral nucleotide. This result holds approximately with other reversible substitution matrices. However, under a RAS model, w is positive (Lockhart et al. 1998). For example, if a site varies in one group under a RAS model, it suggests that the site is evolving rapidly. This in turn suggests that the probability that the site is variable in the second group is higher than it would be with no information about its variability in the first group. In other words, site variation in the two groups is positively correlated. If sites are evolving under the Tuffley and Steel (1998) covariation model, w should also be greater than 0. For example, if a site varies in the first group, it suggests that the site is ON at the parent node of the first group. The probability that the site is ON at the parent node of the second group would be higher than if the site were OFF in the first group. It follows that the probability that the site varies in the second group is also higher than it would be without observing variation in the first group. However, the value of w under a COV + RAS model should be lower than it is under the RAS model. Whereas in the RAS model all fast or slow sites in one group are also fast or slow in the other group, under the covariation model, some of the sites will switch from ON to OFF or OFF to ON between groups. These switches diminish the correlation between group variability. Therefore, the test statistic W is an estimate of a value w that is approximately 0 under a homogeneous (no RAS nor COV) pattern of evolution, positive under a COV or COV + RAS pattern of evolution, and usually even larger under a RAS pattern of evolution. Still, it is not simple to distinguish between models based solely on the value of W .

We can compute w analytically for the different models in a simple example. In the calculations that follow, we assume a simpler RAS distribution and substitution model than in the remainder of the paper. They provide formulae that illustrate why W is a useful statistic for examining the covariation and RAS models. Assume that a rooted tree is formed by two clades (groups) each containing two taxa, that each clade's root node is at distance T from the root

of the tree, and that each taxon is at distance t from the root node of the clade (fig. 1B). Under a Jukes and Cantor (1969) model with a homogeneous rate equal to r , w is 0. Under a RAS model with a proportion of invariable sites (p_{inv}) and the remaining sites evolving at rate r ,

$$w_{\text{ras}} = (9/16)p_{\text{inv}}(1 - p_{\text{inv}})(1 - e^{-8rt/3})^2 \\ \sim 4p_{\text{inv}}(1 - p_{\text{inv}})r^2t^2$$

when t is small. If the evolution process is a COV model, then we can derive from Tuffley and Steel (1998) that

$$w_{\text{cov}} \sim 4\sigma(1 - \sigma)e^{-vT(1-\sigma)}r^2t^2$$

when t is small. We recover the same result as before in case the switching rate v is 0, because the COV model becomes a mixture of invariable sites and constant rate sites with $p_{\text{inv}} = 1 - \sigma$. If we assume a RAS + COV model with the previous parameters, a proportion p_{inv} of sites are permanently invariable sites and a proportion $(1 - p_{\text{inv}})$ of sites can turn ON and OFF. Among the sites that may be variable, a proportion $(1 - p_{\text{inv}})\sigma$ are ON (variable), and a proportion $(1 - p_{\text{inv}})(1 - \sigma)$ are OFF (temporarily invariable), and

$$w_{\text{ras} + \text{cov}} \sim 4(1 - p_{\text{inv}})r^2t^2\{p_{\text{inv}}\sigma^2 + \sigma(1 - \sigma)e^{-vT/(1-\sigma)}\}$$

when t is small. These results first show that w is positive in all of the heterogeneous (RAS, COV, or RAS + COV) models. Second, it shows the conditions under which w values for different models are close or far apart. For instance, w_{cov} in the COV model is further from 0 (homogeneous case) when T is small and t is large, and close to 0 when T is large or t small. The effect of T is exponential, unlike t . Thus, T appears to have a greater effect on w than the t values. It is easier to distinguish the homogeneous and COV models when the clade depth is large and the interior branches are small. The difference in w between the RAS and RAS + COV models is

$$w_{\text{ras}} - w_{\text{ras} + \text{cov}} \sim 4(1 - p_{\text{inv}})r^2t^2\{p_{\text{inv}}(1 - \sigma^2) \\ - \sigma(1 - \sigma)e^{-vT/(1-\sigma)}\}.$$

Therefore, it is usually positive because of the small value of the exponential term, yielding a greater value of w in the RAS model than in the COV + RAS model. As in the previous case, w_{ras} and $w_{\text{ras} + \text{cov}}$ are furthest from each other when the clade depth t is large. However, a small value of T decreases the difference between w_{ras} and $w_{\text{ras} + \text{cov}}$. Thus, a long branch between clades is better to distinguish between the RAS and COV + RAS models.

We present two tests that consider a null hypothesis nested within an alternate covarion hypothesis. In the first test, the *heterogeneity test*, the null hypothesis is a homogeneous model and the alternate hypothesis is any heterogeneous model (COV, RAS, or COV + RAS). If an RAS model has been rejected, this test can use a COV alternate model. In the second test, the *covarion test*, the null hypothesis is the RAS model, and the alternate hypothesis is the COV + RAS model. The null RAS model has gamma-distributed rate heterogeneity (Yang 1994). The distribution of the test statistic W for both tests is obtained using parametric bootstrapping (e.g., Huelsenbeck et al. 1996). The

maximum likelihood parameters used in the parametric bootstrapping are estimated under the null (non-covarion) hypothesis, and many data sets with the same number of characters as the original matrix are simulated using these parameter values. W is computed for each of the simulated data sets to determine its distribution under the null hypothesis. In the heterogeneity test, the P value is the percentage of W 's from simulated data sets that are larger or equal to the observed W , and in the covarion test the P value is the percentage of W 's from simulated data sets that are smaller or equal to the observed W . The calculation of W is implemented in a C program that is available at <http://ginger.ucdavis.edu/>.

The heterogeneity and covarion tests are similar to the previous nonparametric covarion test of Lockhart et al. (1998) but differ in two ways. First, the Lockhart test incorporates an estimate of p_{inv} , whereas the heterogeneity and covarion tests do not. It may be difficult to estimate both p_{inv} and the α shape parameter accurately from a data set with limited taxon sampling (Sullivan et al. 1999). Also, p_{inv} is difficult to estimate when it differs among groups (Steel et al. 2000). Thus, the test of Lockhart et al. (2000) does not require estimation of p_{inv} . The original Lockhart et al. (1998) test statistic together with parametric bootstrapping performed poorly in a preliminary simulation analysis (data not shown). However, the performance of the test greatly improved when p_{inv} was not estimated, and the new test statistic appears to perform well even when data is simulated with a model that includes invariable sites (see below). Second, our test modifies the confidence region $w > 0$ used by Lockhart et al. (1998) because, as demonstrated analytically in the previous example, w should exceed 0 if sites are evolving under the Tuffley and Steel (1998) covarion model. Instead, the significance of the heterogeneity or covarion tests is determined based on the distribution of the test statistic W from parametric bootstrapping.

Performance of the Covarion Tests

The performance of the heterogeneity and covarion tests was examined with simulation experiments. Trees were generated using a conditional pure-birth (Yule) diversification process with a speciation rate of 2.0 using r8s (Sanderson 2003). The model trees were all rooted with a length of 1 substitution per site from the root to tips (fig. 1C and D). To examine the type I error rate (the level) of the covarion test, or how often the test rejects the null hypothesis when it is true, sequences were simulated under an HKY substitution model (Hasegawa, Kishino, and Yano 1985) with a transition to transversion ratio of two, equal nucleotide frequencies, and gamma-distributed rate heterogeneity with four discrete rate categories and a shape parameter $\alpha = 0.25$ (Yang 1994). To examine the power of the covarion test or how often it correctly rejects the null hypothesis, we simulated data sets according to the same model with an additional covarion process. Similarly, the level of the heterogeneity test was determined by simulating characters using an HKY model without rate heterogeneity, and the power was determined by using a COV model and the HKY substitution model described above. For each set of parameter values, 1,000 trees were generated. Each tree

Table 1
GenBank Accession Numbers for the Complete Plastid Genome Sequences Used in the Tests of Covariation Evolution

Taxon	Accession
<i>Adiantum capillus-veneris</i>	NC 004766
<i>Amborella trichopoda</i>	NC 005086
<i>Anthoceros formosae</i>	NC 004543
<i>Arabidopsis thaliana</i>	NC 000932
<i>Atropa belladonna</i>	NC 004561
<i>Calycanthus floridus</i>	NC 004993
<i>Chaetochaeridium globosum</i>	NC 004115
<i>Chlorella vulgaris</i>	NC 001865
<i>Epifagus virginiana</i>	NC 001568
<i>Lotus corniculatus</i>	NC 002694
<i>Marchantia polymorpha</i>	NC 001319
<i>Mesostigma viride</i>	NC 002186
<i>Nephrolepis olivacea</i>	NC 000927
<i>Nicotiana tabacum</i>	NC 001879
<i>Oenothera elata</i>	NC 002693
<i>Oryza sativa</i>	NC 001320
<i>Physcomitrella patens</i>	NC 005087
<i>Pinus koraiensis</i>	NC 004677
<i>Pinus thunbergii</i>	NC 001631
<i>Psilotum nudum</i>	NC 003386
<i>Spinacea oleracea</i>	NC 002202
<i>Triticum aestivum</i>	NC 002762
<i>Zea mays</i>	NC 001666

was used to simulate one sequence alignment, and each simulated data matrix had a length of 1,000 nucleotides.

All simulations were done using a version of Seq-Gen (Rambaut and Grassly 1997) modified to incorporate covariation evolution. The modified version of Seq-Gen is available at <http://www.ginger.ucdavis.edu/>. The test statistic W was calculated using a C program, and maximum likelihood parameter estimates for parametric bootstrapping were obtained using the model tree from PAUP* 4.0b10 (Swofford 2002). The relevant test was then conducted based on 1,000 parametric bootstrap replicates, accepting P values at the 5% level.

We varied several conditions to evaluate sensitivity of the tests. First, the effect of the number of taxa per group on the level and power of the tests were examined. In this experiment, trees had from 4 to 32 taxa contained in two clades of equal size, such that each clade contained 2, 4, 6, 8, or 16 taxa. Additionally, there were two *clade depth* treatments. In the first, the length from the root of the clade to the tips was 0.5 substitutions per site, and in the second it was 0.1 (fig. 1C and D). Thus, the tree depth to clades depth ratio was either 2 (*deep clades*; fig. 1C) or 10 (*shallow clades*; fig. 1D). In order to keep the tree depth equal to 1, branch lengths on the tree were rescaled by a factor of $1/\sigma$ before the simulations in the covariation model simulations. For both of these experiments, the covariation parameters ν and σ were set to 0.4 and 0.6, respectively, as these values are close to estimates in some genes from Huelsenbeck (2002).

We also determined the effect of the RAS gamma distribution parameters on the level and the power of the covariation tests. Simulations were performed as previously described, except that all trees had eight taxa in both clades. To examine the effect of the α shape parameter on the test, simulations were performed with α values of 0.25, 0.5, 1.0, 1.5, and 2.0 with no invariable sites. We also examined the

effect of invariable sites, although the test does not assume their presence. Sequences were simulated with $\alpha = 0.25$ and 0%, 10%, 15%, and 20% invariable sites. These simulations assess the tests' performance using an incorrect model.

Finally, we examined the effect of ν and σ on the power of the heterogeneity and covariation tests. To examine the effect of ν , the simulations were performed as described in the first simulation experiment except that all trees contained eight taxa in both clades, while the value of ν varied from 0.005 to 2 and $\sigma = 0.6$. To examine the effect of σ , ν was set to 0.4 and σ varied from 0.001 to 1.

Covariation Evolution in Plastid Genes

We obtained the amino acid and nucleotide sequences from the protein-coding genes in 23 completely sequenced plant plastid genomes available from GenBank (<http://www.ncbi.nlm.nih.gov/>; table 1). The number of genes within plastid genomes varied from 25 (*Epifagus virginiana*) to 174 (*Chlorella vulgaris*). Sets of potentially homologous amino acid sequences ("clusters") were identified using BLASTCLUST (Dondoshansky 2002) with a clustering threshold of 50% similarity. Of the 72 resulting clusters with at least 12 taxa, we retained 57 that contained at most one sequence from any single taxon. The amino acid sequences from these genes were aligned using ClustalW (Thompson et al. 1994), and the aligned amino acid sequences were used to make nucleotide sequence alignments. The nucleotide alignments are available at <http://ginger.ucdavis.edu/>. All tests used nucleotide sequence data. For all the model tests, we used a reference tree that represents the likely relationships of plant taxa with complete chloroplast sequences (e.g., Pryer et al. 2002; fig. 2), though there is some controversy regarding the phylogeny of the green plant plastid genome (e.g., Goremykin et al. 2003, 2004; D. E. Soltis and P. S. Soltis 2004).

We first performed a likelihood ratio test on each of the 57 genes to evaluate the null hypothesis of equal rates across sites versus the alternate hypothesis of gamma-distributed rate heterogeneity. We calculated the likelihood for each gene using the HKY model and the HKY model with gamma-distributed rate heterogeneity (Hasegawa et al. 1985; Yang 1994). The likelihood ratio test statistic was evaluated based on χ^2 distribution with a 50:50 mixture of 0 and 1 degree of freedom (Self and Liang 1987; Goldman and Whelan 2000; Ota et al. 2000).

Next, we performed the heterogeneity and covariation tests to examine the null hypotheses of homogeneous and RAS evolution for each of the 57 chloroplast genes. Sites with gaps in the alignment may affect the calculation of W , and therefore, W was only calculated from sites that contained no gaps. For the heterogeneity test, the distribution of the test statistic was determined from 2,000 simulations using the maximum likelihood parameters estimated with the HKY model (Hasegawa et al. 1985), with empirical base frequencies and estimating a parameter for ratio of transitions to transversions. Though the parameters were estimated from the full sequence alignment, the simulated data matrices were only as long as the number of sites without gaps (table 2). For the covariation test, the

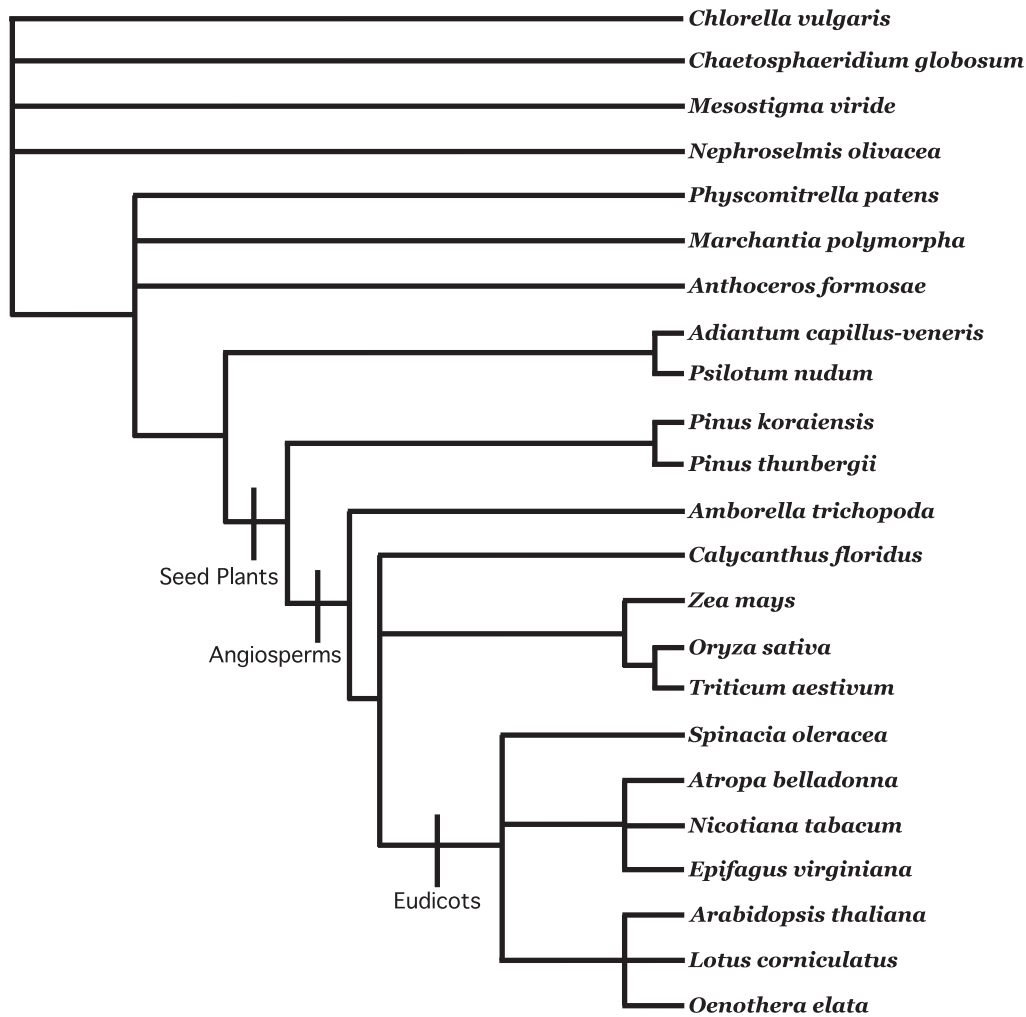


FIG. 2.—Reference tree of taxa with completely sequenced plastid genomes (e.g., Pryer et al. 2002). This topology was used in the tests for covarion evolution along with the MP and ML trees made from a single concatenated alignment of all 57 plastid genes.

distribution of the test statistic was determined from 2,000 data sets simulated using an HKY model with rate variation among sites following a gamma distribution with four discrete rate categories (Yang 1994) and the reference tree (fig. 2). The original groups used to calculate W were the angiosperms and the nonangiosperms (fig. 2). We also performed the covarion test using eudicots and noneudicots groups as well as seed plant and nonseed plant groups (fig. 2). In order to examine the effect of tree topology, we performed the covarion test with the angiosperm and nonangiosperm partition using parsimony and maximum likelihood trees made using the concatenated nucleotide alignment of all 57 genes. The parsimony tree was found using a heuristic tree search starting from a random sequence addition tree with tree bisection reconnection (TBR) branch swapping. The maximum likelihood tree was constructed with the HKY model (Hasegawa et al. 1985) using empirical base frequencies, a transition/transversion ratio of 2, and no rate variability across sites, using the same tree search heuristic as parsimony. Parsimony and likelihood analyses were done using PAUP* (Swofford 2002). We also investigated the presence of covarion evolution at different codon posi-

tions performing all tests on data sets that included the first and second codon positions and data sets that included only the third codon position.

Results

Performance of Covarion Tests in Simulation

The heterogeneity and covarion tests appear to perform well under a wide range of simulation conditions. The level of the tests is close to or below the targeted level of 5% regardless of the number of taxa in each group (fig. 3A). The level appears slightly lower in the covarion test than the heterogeneity test, and the largest difference between the actual and targeted level occurs for the covarion test with shallow clades and few taxa in each clade (fig. 3A). In this case, the test tends to be conservative (level of 0.6% instead of 5%). As the number of taxa increases, the level generally gets closer to the targeted value 5%. The level of the test also remains near 5% in the covarion test as the shape parameter α varies (fig. 4A). However, the level of the test exceeds 5% when there

Table 2
Results of Two Tests for 57 Chloroplast Protein-Coding DNA Sequence Datasets

Gene	Taxa	Aligned bp	Rejects Homogeneity			Rejects RAS-Only		
			123	12	3	123	12	3
<i>atpA</i>	21(10)	1542(1479)	***	***	***		***	
<i>atpB</i>	22(11)	1512(1434)	***	***	***		***	
<i>atpE</i>	22(11)	426(393)	***	***	***		*	
<i>atpF</i>	19(11)	582(522)	***	***				
<i>atpH</i>	22(11)	246(243)	***	***	*			
<i>atpI</i>	22(11)	777(699)	***	***	***		*	
<i>clpP</i>	21(11)	651(585)	***	***	*			
<i>infA</i>	13(3)	246(222)	***	*	***			
<i>ndhA</i>	19(11)	1122(1047)	***	***	***	*	***	
<i>ndhC</i>	19(11)	363(360)	***	***	***			
<i>ndhD</i>	17(10)	1581(1347)	***	***	***	***	***	
<i>ndhE</i>	19(11)	321(276)	***	***				
<i>ndhG</i>	16(11)	609(522)	***	***		***		**
<i>ndhH</i>	19(11)	1182(1173)	***	***	***		***	
<i>ndhI</i>	19(11)	555(468)	***	***		*	***	
<i>ndhJ</i>	18(11)	522(474)	***	***	***			
<i>petA</i>	22(11)	1032(924)	***	***	*		***	
<i>petD</i>	22(11)	567(480)	***	***	***	***	*	
<i>petG</i>	21(11)	114(111)	***	*				
<i>petN</i>	15(11)	87(87)	***		*			
<i>psaA</i>	22(11)	2262(2241)	***	***	***		***	
<i>psaB</i>	22(11)	2205(2202)	***	***	***	***	***	
<i>psaC</i>	22(11)	243(243)	***	***	***			
<i>psaI</i>	19(11)	111(90)	***				*	
<i>psaJ</i>	21(11)	132(123)	***	***	***			
<i>psbB</i>	22(11)	1536(1524)	***	***	***		*	
<i>psbC</i>	22(11)	1461(1383)	***	***	***	*	***	*
<i>psbD</i>	22(11)	1062(1056)	***	***	***		***	
<i>psbE</i>	22(11)	249(243)	***		*		*	
<i>psbF</i>	22(11)	123(117)	***	*	*			
<i>psbH</i>	21(11)	261(219)	***	***	***		***	
<i>psbI</i>	18(10)	114(108)	***					
<i>psbJ</i>	22(11)	126(120)	***	*				
<i>psbK</i>	19(11)	183(159)	***	***	*			
<i>psbL</i>	21(11)	114(114)	***	*	*			
<i>psbM</i>	19(11)	108(102)	***	***				
<i>psbN</i>	22(11)	132(129)	***	*				
<i>psbT</i>	17(11)	114(96)	***	***	*			
<i>psbZ</i>	21(11)	186(186)	***		*	*	***	
<i>rpl14</i>	22(11)	372(360)	***	***	*	*	*	
<i>rpl16</i>	23(12)	429(402)	***	***	***			
<i>rpl20</i>	23(12)	402(336)	***	***	***			
<i>rpl33</i>	21(12)	213(180)	***	***				
<i>rpl36</i>	22(11)	114(111)	***	*	*			
<i>rpoA</i>	18(11)	1119(960)	***	***				
<i>rpoB</i>	21(11)	3471(3042)	***	***	***		*	
<i>rpoC1</i>	20(11)	2232(1914)	***	***	***	***	*	
<i>rpoC2</i>	12(8)	4500(3654)	***	***	***	***	***	
<i>rps11</i>	23(12)	438(381)	***	***	***		***	
<i>rps14</i>	23(12)	309(294)	***	***	***			
<i>rps16</i>	14(11)	264(165)	*	*		*		
<i>rps2</i>	19(11)	717(678)	***	***	***	*		
<i>rps3</i>	22(12)	858(609)	***	***	*			
<i>rps4</i>	22(12)	741(540)	***	***	***			
<i>rps8</i>	23(12)	438(381)	***	***	*			
<i>ycf3</i>	18(9)	549(468)	***	*	*		*	
<i>ycf4</i>	20(10)	609(549)	***	***	*		*	
total			57	52	44	14	26	2

NOTE.—The first test compares a heterogeneous model (RAS and/or COV) to a null model of homogeneity. The second test compares the RAS+COV model to the null RAS model. Gene names are given as standard GenBank gene labels. For each gene, the total taxa are the number of taxa out of 23 that had the gene, and the number of angiosperm taxa are in parentheses. The aligned bp column shows the length of the alignment for each gene followed by the number of sites without gaps in parentheses. Numbers (123, 12, 3) indicate codon position(s) included used in the test. *** $P < 0.0005$; ** $P < 0.005$; * $P < 0.05$.

are deep clades and the proportion of invariable sites is 15% or 20% (fig. 4B).

The tests' power increases with the number of taxa per clade (fig. 3B). The power of the test is often higher

with deep clades and a short intermediate branch than with shallow clades and a long intermediate branch (figs. 3B, 4, and 5). Power is low with only two taxa in each clade, except in case of the heterogeneity test with deep

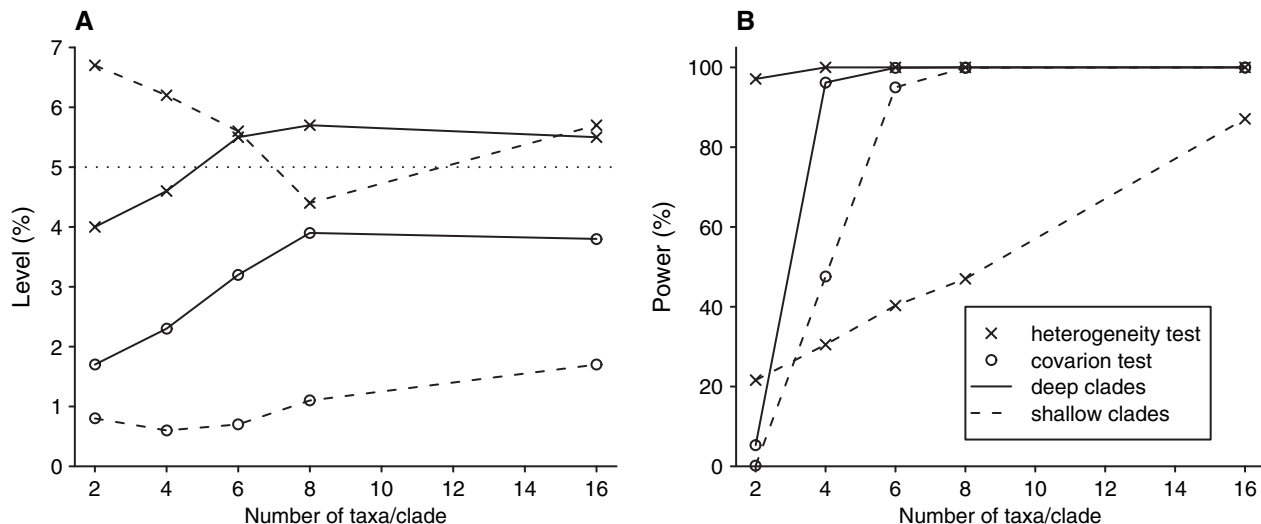


FIG. 3.—Results of the simulation study testing the effect of group size on the level and power of the heterogeneity and covarion tests. Each point on the graphs represents 1,000 simulated data sets. The covarion test was performed on data sets simulated with a RAS model using shape parameter $\alpha = 0.25$, and the heterogeneity test was performed on data sets simulated without rate variation across sites. In figure 1A, the simulations were performed under the null model, and the level of the tests is the percentage of simulation replicates in which the null hypothesis was mistakenly rejected. In figure 1B, the simulations were performed under a covarion model (COV only for the heterogeneity tests; RAS + COV for the covarion test). The power of the test shows the percentage of simulation replicates that correctly rejected the null hypothesis.

clades (fig. 3B). Six taxa in each clade are usually needed to reach a power of 90%, except for the heterogeneity test with shallow clades, which requires greater than 16 taxa in each clade (fig. 3B). The power of the covarion test is little affected by α or the proportion of invariable sites (fig. 5A). The test power is still above 90% when α is 2 as well as when the proportion of invariable sites is 0.2 (fig. 4). In the covarion test, the power is high when σ is between 0.04 and 0.7 and when ν is below 1. The latter condition should be fulfilled on real data, as the switching process is expected to be much slower than the substitution process. The range of good performance is smaller for the heterogeneity test, requiring σ between 0.2 and 0.6 and ν below 0.6 switches/substitution.

Plastid Genes

The likelihood ratio test strongly rejects ($P < 0.0005$) a homogenous rates model in favor of a RAS model in all 57 genes for all sites and for first and second codon positions only. The homogenous rates model was also strongly rejected in 56 of the 57 genes when only third codon positions were included in the test, the exception being *petN* ($P = 0.115$).

The heterogeneity test detected significant heterogeneity (RAS and/or COV) in all of the 57 data sets when all codon positions are included (table 2). When applied to first and second positions only, heterogeneity was detected in 52 out of 57 genes, and in 44 of 57 genes when only third positions are included (table 2). The covarion test detected

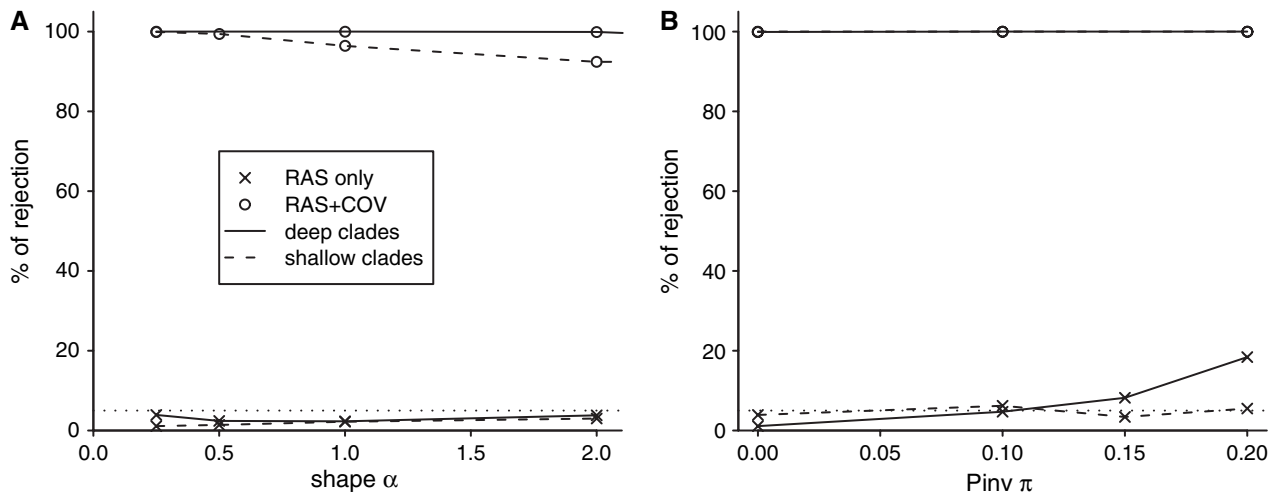


FIG. 4.—Results of the simulation study showing the effect of the α shape parameter and p_{inv} on the covarion test. When simulations use a covarion model (circles), the rejection percentage is the power of the test. When simulations do not use a covarion model (indicated by x in the graph), the rejection percentage is the level of the test. Rate variation across sites was simulated as a mixture of discrete gamma distribution with invariable sites. The tests assume $p_{inv} = 0$. In the simulations for (A), the α shape parameter for the gamma distribution varied and $p_{inv} = 0$. In the simulations for (B), $\alpha = 0.25$ and p_{inv} varied.

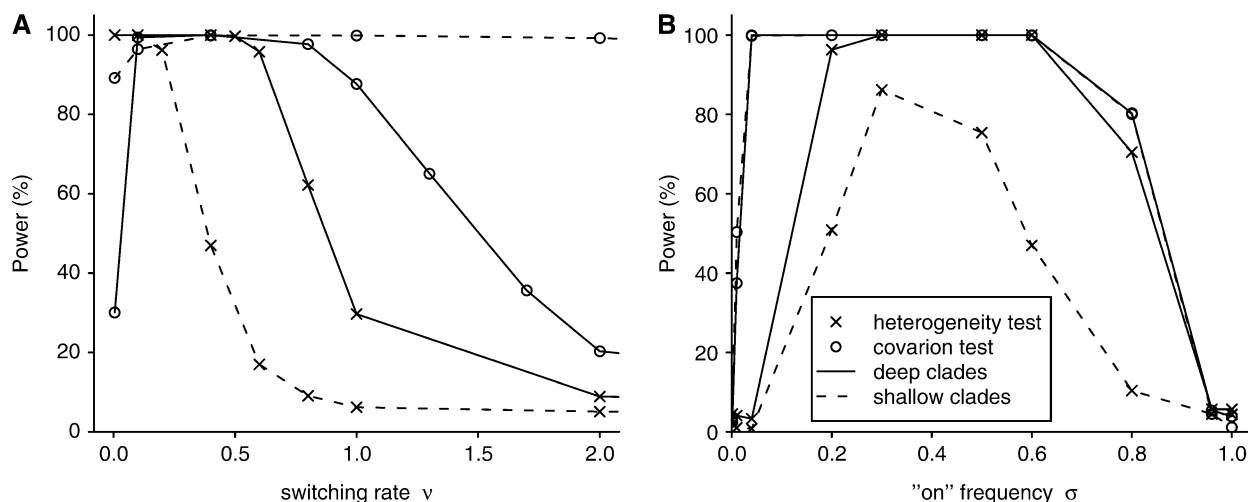


FIG. 5.—Results of the simulation study showing the effect of the covariation parameters (ν and σ) on the power of the heterogeneity and covariation tests. In the simulations for (A), ν is varied and $\sigma = 0.6$. In the simulations for (B), σ is varied, $\nu = 0.4$ switches/substitution.

covariation evolution in 14 of 57 genes across all positions; however, covariation evolution is detected in 26 out of 57 genes when only first and second positions are included (table 2). Only two genes show evidence of a covariation structure in third codon positions, which is not greater than the number of significant tests we would expect by chance alone (table 2).

The results of the covariation test are nearly the same when we used the maximum parsimony (MP) tree and maximum likelihood (ML) tree from the combined data matrix instead of the reference tree (table 3). All of the 57 genes still rejected the homogeneity model with all positions included. With first and second codon positions, heterogeneity was detected in 53 (MP tree) or 52 (ML tree) data sets, and in 42 data sets (for both MP and ML trees) for third

positions (table 3). Covariation evolution was detected in 14 (MP tree) or 16 (ML tree) genes when all positions were included, 24 genes with first and second positions, and two genes with third positions only (table 3).

Fewer significant results for the covariation test were obtained when taxa were partitioned into eudicot and non-eudicot groups rather than angiosperm and nonangiosperm groups, but slightly more significant results were obtained using the seed plant and nonseed plant groups (table 3). In the covariation test with eudicot and noneudicot groups, the RAS-only model was rejected in only one or two genes with all codon positions, 11 genes (7 for MP tree, 10 for ML tree) with first and second position and two genes with third positions only (table 3). However, the covariation test with the seed plant–nonseed plant partition generally rejected the RAS model at least as many times as with the angiosperm–nonangiosperm groups. With the seed plant partition, the covariation test rejected the RAS model 15 or 18 times (depending on the tree) for all sites and 27 or 32 times for only the first and second codon positions (table 3). The results of the heterogeneity test were nearly uniform for all sets of groups. Either 56 or all 57 genes rejected the homogeneity model with all positions included, and 51 or 53 genes rejected homogeneity with first and second positions (table 3).

Table 3
Summary of the Results from Heterogeneity and Covariation Tests

Tree	Group	Rejects Homogeneity			Rejects RAS-Only		
		123	12	3	123	12	3
Reference	Angiosperm	57	52	44	14	26	2
MP	Angiosperm	57	53	42	14	24	2
ML	Angiosperm	57	52	42	16	24	2
Reference	Eudicot	57	52	41	1	11	2
MP	Eudicot	57	53	42	2	7	2
ML	Eudicot	57	52	41	2	10	2
Reference	Seed plant	56	52	39	18	32	5
MP	Seed plant	56	51	38	15	27	4
ML	Seed plant	56	52	38	15	27	4

NOTE.—The heterogeneity test (left columns) evaluates the null hypothesis of homogeneous evolution among sites, and the covariation test (right columns) evaluates the null hypothesis of RAS evolution. Tree refers to topology used in the covariation test, either the reference tree (fig. 1), or the maximum parsimony (MP) or maximum likelihood (ML) tree made from the nucleotide alignment of all 57 loci. Group represents the bipartition used to calculate the test statistic. The three bipartitions separated angiosperms and nonangiosperms, eudicots and noneudicots, and seed plants and nonseed plants. Numbers (123, 12, 3) indicate codon position(s) included in the analysis. The numbers in the table are the significant tests ($P < 0.05$) from the 57 individual gene tests.

Discussion

Though the idea of covariation evolution was proposed over 30 years ago (Fitch and Markowitz 1970; Fitch 1971), there is still surprisingly little data regarding its frequency or importance. Huelsenbeck (2002) rejected a noncovariation model in favor of a covariation model for 9 out of 11 loci from a variety of organisms, but few other studies have examined more than one or two loci for evidence of covariation evolution (e.g., Lockhart et al. 1998; Galtier 2001; Misof et al. 2002). We present a large-scale analysis of covariation evolution in which we examined most genes from completely sequenced green plant plastid genomes. Our test for covariation patterns of evolution performs well under a wide

range of simulated conditions (figs. 3–5) and is simple to implement. Using this test, nearly half of the plastid genes show evidence of covarion evolution in the first two codon positions (table 2), indicating that changes in selective constraints of amino acids through time are an important factor in creating the sequence variation among many plastid genes.

The simulations and the analysis of the plastid genes indicate several conditions under which the heterogeneity and covarion tests will perform well. The level of the tests are near the desired 5%, except in the heterogeneity test when the interior branch separating the two groups is short and the terminal branches are long or when there is a high percentage of invariable sites (figs. 3A and 5). Thus, the frequency of type 1 error should be relatively low using the covarion test or when using the heterogeneity test when the interior branch separating the groups is long compared to the terminal branches. The number of taxa per group also is clearly important for the power of the covarion test, and we recommend that groups have at least six taxa (fig. 3B).

The choice of the bipartition can greatly affect the performance of the heterogeneity and covarion tests. For example, there were many fewer significant covarion tests of the plastid genes using the eudicot–noneudicot bipartition than the angiosperm–nonangiosperm bipartition (table 3). The heterogeneity and covarion tests should work using any true bipartition of the gene phylogeny but the choice of bipartition can affect the power of the test. We suggest selecting a bipartition based on tree properties, such as the group sizes and the length of the interior branch separating the groups. The interior branch should not be too short so there can be switches in the ON or OFF state of a site between groups. These switches will diminish the correlation of site variability among groups due to the RAS model. Furthermore, the bipartition must also leave an adequate number of taxa in each group. The tree topology can also affect the performance of the test. If the topology is wrong, a significant result may reflect a rejection of the topology rather than the stated null hypothesis. In the plastid data, the covarion test performs similarly using the maximum likelihood or parsimony topologies inferred using all 57 loci (table 3). The results were also similar using the ML or MP topologies of the individual genes (data not shown). However, we suggest using only well-supported topologies to obtain the most accurate results.

As expected, the power of the tests decreases when covarion parameters are so extreme that they cause the covarion model to resemble the null hypothesis. Power decreases for both the heterogeneity and covarion tests when v is large. For example, the power of the covarion test is 80% when $v = 5$ and drops further as v increases (data not shown), and the heterogeneity test appears to be even more sensitive to v . Power also decreases when $v = 0$ for the covarion test (fig. 5A). However, it is not the case when $v = 0$ in the heterogeneity test because when $v = 0$, the sequences are a heterogeneous mixture of invariable and constant rate sites. The power of the heterogeneity and covarion tests also drops when the ON frequency σ is zero or one (fig. 5B). The power remains high for the low values of σ due to rescaling the branch lengths by a factor of $1/\sigma$, which is necessary to keep an average of one substi-

tution from root to tip of the trees. Due to this constraint, the evolution model does not converge to completely invariable sites when σ approaches zero. However, the model slowly converges to the null hypothesis model (either a homogeneous or RAS model) as σ nears 0.

Other tests of covarion evolution are computationally complex and have seen little use (Galtier 2001; Huelsenbeck 2002). The likelihood ratio test for gamma-distributed rate variation across sites appears to be more powerful than our heterogeneity test (table 2). Kelly and Rice (1996) proposed another likelihood ratio test for any distribution of rate variation across sites that use parametric bootstrapping. If variation in the rate of evolution across sites is causing the heterogeneous patterns of evolution, it may be best to test for heterogeneity with a likelihood ratio test before using the covarion test. However, a likelihood ratio test may not detect heterogeneity if it is caused by covarion evolution without significant rate variation across sites. Therefore, if a likelihood ratio test fails to detect significant heterogeneous evolution, we suggest using our heterogeneity test to see if the COV model may be appropriate. In the plastid loci, both the likelihood ratio test for variation in rates across sites and the heterogeneity tests were nearly always significant, suggesting that heterogeneity in the process of evolution is nearly ubiquitous. Thus, the covarion test is likely more important than the heterogeneity test. The covarion test is a fast and computationally simple alternative to likelihood ratio tests, and it should be useful for screening large numbers of loci for covarion evolution. Also, unlike the current likelihood ratio tests, its performance has been extensively tested with simulations (figs. 3–5). The covarion test detected evidence of covarion drift in 26 plastid loci (tables 2 and 3), and there is strong evidence that this is due to evolution in the rate of nonsynonymous substitution. The most significant tests of covarion evolution were from data sets that included only the first and second codon positions, which are much more likely to represent nonsynonymous substitutions than the third codon position (tables 2 and 3). Thus, adding the third codon position sites appears to mask evidence of covarion drift, and there is no evidence of significant change in selective constraints of third codon positions. The increased frequency of significant covarion tests using only first and second codon positions is consistent with the hypothesis that changes in selective constraints of the amino acids play an important role in the evolution of the observed patterns of sequence variation in many plastid genes. The size of the plastid gene data sets, especially when only some of the codon sites were included, were often much smaller than the 1,000-bp simulated data sets used to examine the power of the tests suggesting that the signal for covarion evolution is strong.

The apparent prevalence of covarion patterns of evolution further suggests that models that do not incorporate rate variation through time may not be adequate for evolutionary inference. The apparent general lack of covarion structure in third codon positions also suggests that the different codon positions evolve under different processes and may not be described adequately with a single model of evolution. We note that our test assumes a stationary covarion drift model of evolution, in which the sites change

from the ON to OFF state throughout all lineages in the tree. It does not explicitly test for covariation shift in which there is a large change in the proportion of invariant sites in a specific lineage or part of the tree. It is thus possible that our test is not detecting all instances of covariation evolution. Previous studies have argued that failing to account for the rate variation of sites through time may be problematic for inferring phylogenies, and a covariation structure might help explain the presence of a phylogenetic signal among anciently diverged lineages (Lockhart et al. 1998, 2000; Lopez et al. 1999; Philippe and Germot 2000; Steel et al. 2000). Still, few phylogenetic studies have implemented covariation models of evolution. Vogl et al. (2003) demonstrated that, using standard models of evolution, many plastid genes appear to have significantly different phylogenies, and they postulated that a covariation process of evolution in some loci may explain some of the incongruence. Given our findings that covariation evolution is in fact commonplace among these genes, it would be interesting to examine whether incongruence was in part due to reconstructing gene trees without taking covariation evolution into account.

Acknowledgments

This research was funded by NSF grant DEB0075319. We thank Peter Lockhart and two anonymous reviewers for helpful comments.

Literature Cited

- Dondoshansky, I. 2002. Blastclust (NCBI Software Development Toolkit). NCBI, Bethesda, Md.
- Fitch, W. M. 1971. Rate of change of concomitantly variable codons. *J. Mol. Evol.* **1**:84–96.
- Fitch, W. M., and E. Markowitz. 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 covariation-like model. *Mol. Biol. Evol.* **18**:866–873.
- Goldman, N., and S. Whelan. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **17**:975–978.
- Goremykin, V. V., K. I. Hirsch-Ernst, S. Wölfl, and F. H. Hellwig. 2003. Analysis of the *Amborella trichopoda* chloroplast genome suggests that *Amborella* is not a basal angiosperm. *Mol. Biol. Evol.* **20**:1499–1505.
- . 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. *Mol. Biol. Evol.* **21**:1445–1454.
- Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **21**:160–174.
- Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**:275–284.
- Huelsenbeck, J. P. 2002. Testing a covariation model of DNA substitution. *Mol. Biol. Evol.* **19**:698–707.
- Huelsenbeck, J. P., D. M. Hillis, and R. Jones. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. Pp. 19–45 in J. D. Ferraris and S. R. Palumbi, eds. *Molecular zoology: advances, strategies, and protocols*. Wiley-Liss, New York.
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 in H. N. Manro, ed. *Mammalian protein metabolism*. Academic Press, New York.
- Kelly, C., and J. Rice. 1996. Modeling nucleotide evolution: a heterogeneous rate analysis. *Math. Biosci.* **133**:85–109.
- Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78**:454–458.
- Liò, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. *Genome Res.* **8**:1233–1244.
- Lockhart, P. J., D. Huson, U. Maier, M. J. Fraunholz, Y. Van de Peer, A. C. Barbrook, C. J. Howe, and M. A. Steel. 2000. How molecules evolve in eubacteria. *Mol. Biol. Evol.* **17**:835–838.
- Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. H. Huson, M. A. Charleston, and C. J. Howe. 1998. A covariation model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* **15**:1183–1188.
- Lopez, P., P. Forterre, and H. Philippe. 1999. The root of the tree of life in the light of the covariation model. *J. Mol. Evol.* **49**:496–508.
- Misof, B., C. L. Anderson, T. R. Buckley, D. Erpenbeck, A. Rickert, and K. Misof. 2002. An empirical analysis of mt 16S rRNA covariation-like evolution in insects: site-specific rate variation is clustered and frequently detected. *J. Mol. Evol.* **56**:330–340.
- Miyamoto, M. M., and W. M. Fitch. 1995. Testing the covariation hypothesis of molecular evolution. *Mol. Biol. Evol.* **12**:503–513.
- Ota, R., P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol. Biol. Evol.* **17**:798–803.
- Penny, D., B. J. McCormish, M. A. Charleston, and M. D. Hendy. 2001. Mathematical elegance with biochemical realism: the covariation model of molecular evolution. *J. Mol. Evol.* **53**:711–723.
- Philippe, H., and A. Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* **17**:830–834.
- Posada, D., and K. A. Crandall. 2001. A comparison of different strategies for selecting models of DNA substitution. *Syst. Biol.* **50**:580–601.
- Pryer, K. M., H. Schneider, E. A. Zimmer, and J. A. Banks. 2002. Deciding among green plants for whole genome studies. *Trends Plant Sci.* **7**:550–554.
- Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte-Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**:235–238.
- Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**:301–302.
- Self, S. G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**:605–610.
- Shoemaker, J. S., and W. M. Fitch. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol. Biol. Evol.* **6**:270–289.
- Soltis, D. E., and P. S. Soltis. 2004. *Amborella* not a “basal angiosperm”? Not so fast. *Am. J. Bot.* **91**:997–1001.
- Steel, M., D. Huson, and P. J. Lockhart. 2000. Invariable sites models and their use in phylogeny reconstruction. *Syst. Biol.* **49**:225–232.

- Sullivan, J., D. L. Swofford, and G. J. P. Naylor. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* **16**:1347–1356.
- Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 10. Sinauer Associates, Sunderland, Mass.
- Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Mable, eds. *Molecular systematics*. Sinauer, Sunderland, Mass.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- Tuffley, C., and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**:63–91.
- Vogl, C., J. Badger, P. Kearney, M. Li, M. Clegg, and T. Jiang. 2003. Probabilistic analysis indicates discordant gene trees in chloroplast evolution. *J. Mol. Evol.* **56**:330–340.
- Whelan, S., P. Liò, and N. Goldman. 2001. Molecular phylogenetics: state of the art methods for looking into the past. *Trends Genet.* **17**:262–272.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites. *J. Mol. Evol.* **39**:306–314.
- . 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* **11**:367–372.

Peter Lockhart, Associate Editor

Accepted December 21, 2004