An Improved Statistical Method for Detecting Heterotachy in Nucleotide Sequences

Guy Baele,*† Jeroen Raes,†[1] Yves Van de Peer,† and Stijn Vansteelandt*

*Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium; and †Department of Plant Systems Biology, Ghent University, Ghent, Belgium

The principle of heterotachy states that the substitution rate of sites in a gene can change through time. In this article, we propose a powerful statistical test to detect sites that evolve according to the process of heterotachy. We apply this test to an alignment of 1289 eukaryotic rRNA molecules to 1) determine how widespread the phenomenon of heterotachy is in ribosomal RNA, 2) to test whether these heterotachous sites are nonrandomly distributed, that is, linked to secondary structure features of ribosomal RNA, and 3) to determine the impact of heterotachous sites on the bootstrap support of monophyletic groupings. Our study revealed that with 21 monophyletic taxa, approximately two-thirds of the sites in the considered set of sequences is heterotachous. Although the detected heterotachous sites do not appear bound to specific structural features of the small subunit rRNA, their presence is shown to have a large beneficial influence on the bootstrap support of monophyletic groups. Using extensive testing, we show that this may not be due to heterotachy itself but merely due to the increased substitution rate at the detected heterotachous sites.

Introduction

It has been extensively shown that the introduction of rates across sites (RAS) models can offer vast improvements in reconstructing phylogenies (Olsen 1987; Van de Peer, Rensing et al. 1996; Yang 1996; Van de Peer et al. 2000). Such models postulate that the substitution rate of a site (i.e., a nucleotide in a nucleic acid sequence) is constant through time (i.e., in all lineages) but allow this rate to vary between sites. This usually happens by letting a gamma distribution express the variability of the substitution rates (Yang 1996), eventually leading to so-called slow- and fast-evolving sites. It has been argued that the site-specific selective constraints that lie at the origin of RAS may also vary in time and between lineages. Indeed, Fitch and Markowitz (1970) observed that the evolutionary rate of a particular site in coding sequences can be variable across the phylogeny due to the fact that sites critical with respect to the function of a macromolecule may change within the nucleotide sequence over time. Specifically, their covarion hypothesis postulates that, at any given time, only a fraction of the sites can accept substitutions. These sites are called concomitantly variable codons (covarions) and their relative occurrence is assumed constant over time. To this end, when a substitution is accepted, it is assumed that another site becomes invariable and vice versa. The underlying biological argument goes that as mutations are fixed at some sites in a gene, the functional constraints at other sites may change.

Recently introduced tests have convincingly confirmed, using real data, that the substitution rate of a site is not always constant through time (Lockhart et al. 1998; Gu 1999; Misof et al. 2002) but did not validate the covarion model as a sufficient explanation of sequence evolution. This is partly because a constant percentage of covarions

in the covarion hypothesis may be overly restrictive (Steel et al. 2000). A related process, called heterotachy, that enables greater generality allows for site-specific rate variation (SSRV) regardless of the possible presence of covarions. Under such process, the evolutionary rate at a site may be different in different parts of the tree (Philippe and Lopez 2001). Specifically, evolutionary rates may change over time for each site separately. Thus, heterotachy is a site property that allows the ratio of substitution rates on different branches of the tree to vary across sites (Lockhart and Steel 2005; Lockhart et al. 2006). One specific form of heterotachy, which assumes that the rate of change between substitution rates is constant over sites, can be modeled by superimposing a continuous rate switching process (Galtier 2001) on Yang's RAS model (Yang 1996) to allow the rate at a given site to vary over time. The model of Galtier (2001) allows SSRV at independent sites.

Evolutionary models that describe the covarion or heterotachy hypothesis may provide a better description of the data than models that do not allow constraints to change over time (Fitch and Markowitz 1970; Fitch 1971; Tuffley and Steel 1998; Huelsenbeck 2002). Indeed, Huelsenbeck (2002) used likelihood ratio tests to show that the covariotide model of Tuffley and Steel (1998) provides a better explanation of evolution at several genes than a model that does not allow rates of substitution to change over time (Huelsenbeck 2002). Further, Lockhart et al. (1996) showed that inference of evolutionary trees under models that do not allow SSRV can be biased in the presence of covarion patterns of change. These results are suggestive of the importance of detecting heterotachous sites in an alignment because the presence of such positions might influence the choice of an evolutionary model (with/ without heterotachy). Furthermore, knowing which sites evolve under time-varying rates could provide important insights into evolutionary processes.

Although tests for unveiling heterotachous sites have been proposed in the past (Lopez et al. 1999), we argue in this article that existing tests are restrictive because 1) they may incorrectly detect many nonheterotachous sites as a result of multiple testing errors and 2) results may be highly sensitive to the number of sites in the sequence and to the number of sequences. To accommodate these problems, we have developed a new statistical test to detect heterotachy,

which corrects for multiple testing by controlling the false discovery rate (FDR) (Benjamini and Hochberg 1995; Storey and Tibshirani 2003). We have applied this test to a large number of eukaryotic small subunit ribosomal RNA (SSU rRNA) sequences and estimated that heterotachy is present in 66% of all sites of our alignment. Identifying which sites are heterotachous is more demanding. Controlling the FDR at 5%, we could identify 29% of all sites as being heterotachous with good confidence. We have used the results to investigate whether the presence of SSRV is related to specific monophyletic groups or to secondary structure features of the SSU rRNA. Further, we have examined the impact of heterotachous sites on the bootstrap support of certain monophyletic groups. As in the study of Lockhart et al. (1998) on covariotide substitution, we observe that the removal of heterotachous sites decreases bootstrap support under evolutionary models that do not acknowledge SSRV. We clarify the possible causes for such a decrease through extensive testing.

Materials and Methods
Data Collection

SSU rRNA sequences were extracted from the European ribosomal RNA database (Wuyts et al. 2004). The extracted sequences were aligned, taking into account the secondary structure information derived by comparative sequence analysis of thousands of sequences. Sites that contained gaps for some monophyletic groups were removed, as well as sequences that could not be aligned properly. Monophyletic groups containing fewer than 15 sequences were removed. This resulted in a data set of 1289 unique sequences with a length of 968 nt divided over 21 monophyletic groups: Acanthamoeba (17), Acanthocephala (21), Annelida (85), Apicomplexa (47), Arthropoda (80), Ascomycota (59), Bacillariophyta (41), Bangiophyceae (33), Basidiomycota (81), Chlorophyta (81), Chordata (60), Ciliophora (75), Cnidaria (76), Cryptomonadaceae (18), Embryophyta (87), Euglenida (48), Florideophyceae (89), Kinetoplastida (53), Mollusca (82), Platyhelminthes (75), and Zygomycota (81).

Test of Heterotachy

To detect heterotachy at a given site, the number of substitutions for each site in each monophyletic group was predicted using a combination of neighbor-joining and maximum likelihood. First, a neighbor-joining tree was calculated using PAUP* (Swofford 2001). Based on this tree, the parameters for the general time-reversible (GTR) evolutionary model with among-site rate variation were estimated using maximum likelihood. Finally, trees were computed for each of the monophyletic groups separately using neighbor joining, using the parameters estimated by maximum likelihood. The substitution rates for each site were calculated with PAML (Yang 1997) and used to predict substitution numbers. As such, a 21 (corresponding to 21 monophyletic groups) by 968 (length of the sequence alignment) matrix of substitution numbers was created. We use O_{ij}, $i = 1, \ldots, n = 21$, $j = 1, \ldots, 968$, to refer to the predicted number of substitutions at site j in group i.

We define a site to be homotachous (i.e., not heterotachous) when the expected number of substitutions at that particular site in each lineage i is proportional to the overall substitution rate λ_i of that lineage, as measured for instance by the tree length of that lineage. To test whether a given position j is heterotachous, we propose the following chi-square test:

$$\chi_j^2 = \sum_{i=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad (1)$$

where

$$E_{ij} = \frac{\lambda_i}{\sum_{k=1}^{n} \lambda_k} \cdot \sum_{k=1}^{n} O_{kj} \qquad (2)$$

is the expected number of substitutions in the ith group under the null hypothesis of homotachy. Equation 2 is obtained on noting that under the null hypothesis, the expected number of substitutions is proportional to the overall substitution rate λ_i in that group. The Test statistic 1 builds on the chi-square test of Lopez et al. (1999) but differs from it in the sense that the tree lengths are not considered as an independent column of data but as constants used for comparison purposes. A desirable consequence is that our test statistic does not change when the overall rates change proportionally, and hence the decision whether a given site is heterotachous is not affected by the number of sites in the alignment (which itself affects the tree length). Such changes would not be allowed because proportional changes of the overall evolutionary rates do not modify the null hypothesis (i.e., the definition of homotachy).

It can be shown that the modified Test statistic 1 follows a chi-square distribution with n degrees of freedom (df) in large samples. Because the asymptotic distribution of the chi-square test is unreliable with low cell values (i.e., substitution numbers), we have chosen to use permutation tests. In each of 100 000 permutations, the substitutions for each site were redistributed over the monophyletic groups in the following way: we kept the tree lengths for the different groups fixed and chose the probabilities of assigning substitutions to a monophyletic group proportional to the average rate (or the tree length) of that group (see Appendix C). The latter assures that the data are generated under homotachy. By comparing the chi-square statistics of the original data set to the chi-square statistics of each of the permutations, a P value is assigned to each site (Roff and Bentzen 1989), indicating the degree of evidence against the presence of homotachy.

Multiple Testing Problem

Because the chi-square test 1 is used for each of the 968 sites in our alignment separately, the overall risk of false detections is high. Although Bonferroni correction can be used to control the risk of at least one false detection over all sites, it aims to control errors in the unrealistic situation where there is no heterotachy at any site. Furthermore, Bonferroni correction tends to be conservative and hence underpowered. We have therefore chosen to control the FDR (Benjamini and Hochberg 1995), which is defined
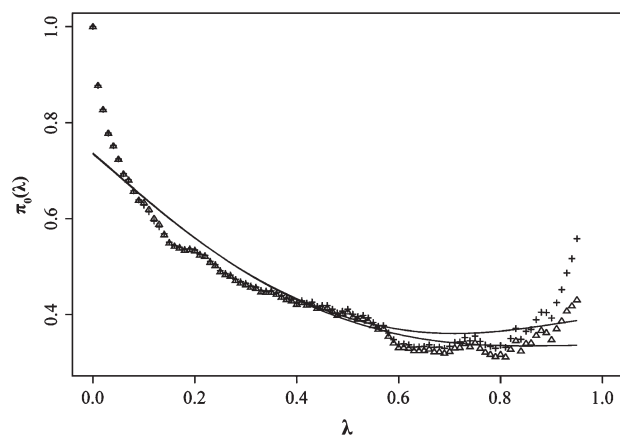
FIG. 1.—Exact distribution of P values decreases conservativeness. Fitted splines before (top spline, "+" symbols) and after (bottom spline, triangular symbols) correcting for nonuniform P values are used to approximate the proportion of null hypotheses ($\pi_0(1)$). Although the overall trend is very similar, the difference at the ending points becomes important. Because a prediction at $\lambda=1$ is required, not correcting for nonuniformity could result in seriously overestimated q values.

in our setting as the proportion of false results among the sites that were detected to be heterotachous.

The FDR can be controlled below 5% by rejecting the null hypothesis at all sites with q value (Storey 2003; Storey and Tibshirani 2003) less than 5%. The latter is the smallest FDR at which the test would still reject and, similar to the P value, expresses the amount of evidence against the null hypothesis (smaller indicating more evidence against the null hypothesis).

Assumption of Uniform P Values

The calculation of q values (Storey and Tibshirani 2003; see Appendix B) requires knowing the proportion of truly homotachous positions (Storey 2003). Storey and Tibshirani (2003) propose to estimate this by calculating, for a range of λ values in $]0, 1[$, the observed number of P values greater than λ divided by the expected number of P values greater than λ under the null hypothesis. Assuming that P values are uniformly distributed under the null hypothesis, this is:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; \ i = 1, \ldots, m\}}{m \cdot (1 - \lambda)}. \tag{3}$$

This equation is then used to approximate the proportion of truly homotachous sites $\hat{\pi}_0(1)$.

Because cell values (i.e., substitution numbers) are low, the assumption that P values are uniformly distributed under the null hypothesis is invalid, and hence, Equation 3 is not applicable. The exact distribution of P values under the null hypothesis therefore needs to be determined. To this end, 2000 permutations of the original data set were calculated using a similar process as explained above. Next, the mean number of P values greater than λ was determined for each site, with λ ranging from 0 to 1 and subsequently substituted in the denominator of $\hat{\pi}_0(\lambda)$. Figure 1 illustrates the effect this has on the estimation of $\hat{\pi}_0(\lambda)$. It can be concluded from figure 1 that not adjusting for the nonuni-

Table 1
Classification of Sites According to Their q Values

	Category (q level)	Positions
Heterotachy	<2.5%	123
	]2.5%, 5%]	160
	]5%, 7.5%]	134
	]7.5%, 10%]	112
	>10%	439

NOTE.—The 968 sites of our alignment are classified into 5 categories, according to their q value. q values lower than 5% indicate evidence in favor of heterotachy (q values lower than 2.5% indicate strong evidence) and q values higher than 5% indicate lack of evidence in favor of heterotachy.

form P values would make our method too conservative and hence underpowered. Figure 1 additionally shows that the proportion of truly heterotachous (homotachous) sites $1 - \hat{\pi}_0(1)(\hat{\pi}_0(1))$ is estimated to be 66.3% (33.7%).

Results

Test of Heterotachy

On a sequence alignment of 1289 eukaryotic SSU rRNA sequences of 968 sites, our method identified 283 (or 29.2%) heterotachous sites (i.e., sites at which the null hypothesis of homotachy is rejected and thus have a q value below 5%) while controlling the FDR at 5%. These sites were subdivided into 5 categories, from strong evidence in favor of heterotachy (q value below 2.5%) to moderate evidence (q value between 2.5% and 5%) (see table 1). The remaining sites, that is, the sites that are not rejected and thus have a q value above 5% will be labeled as homotachous sites below.

As reported above, the test of Lopez et al. (1999) considers the tree lengths as a column of independent data. By doing so, this method identifies different (numbers of) sites as being heterotachous sites depending on whether one compares substitution numbers with the tree length or proportional measures of it. Using the tree length, 366 sites were classified as heterotachous by the method of Lopez et al. (1999), as compared with 378 using the average number of substitutions (i.e., the tree length divided by the number of sites). For other proportional measures, such as 10 or 100 times the average, this method classified 361 and 321 sites as being heterotachous. Results from our test are more reliable because each time the same 283 sites were detected to be heterotachous.

Phylogenetic Groups

After the identification of heterotachous sites in our data set, we determined which phylogenetic groups were primarily responsible for heterotachy at a given site. Therefore, the contribution of each monophyletic group to the chi-square statistic 1 of a site was calculated. To conclude that a given monophyletic group is responsible for heterotachy at a given site, such a contribution must be significantly elevated compared with its expectation under the null hypothesis. We therefore estimated the 95% percentile of each contribution under the null hypothesis using 10 000 permutations. Contributions exceeding this percentile were taken as evidence that heterotachy was caused by the evolutionary rate being unexpectedly high
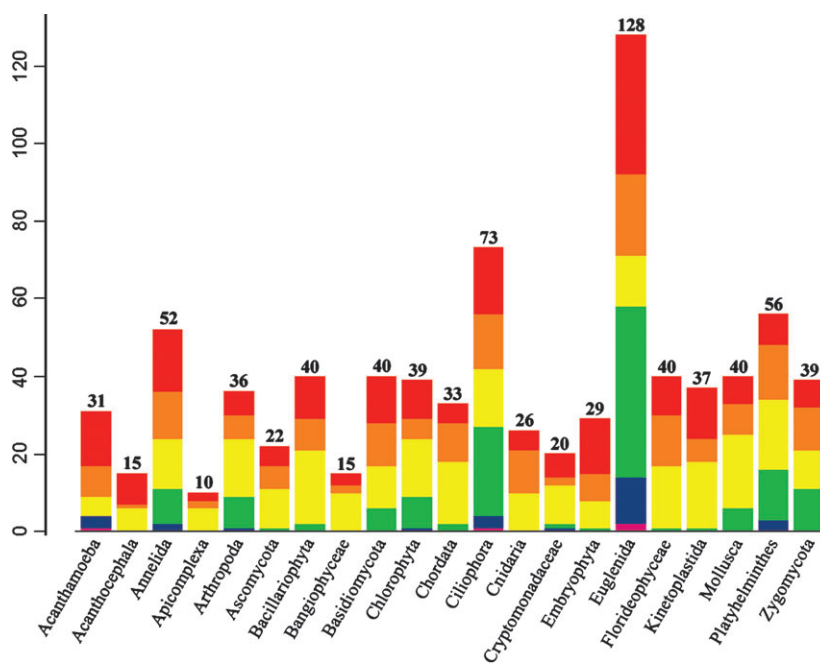
Fig. 2.—Significant contributions to heterotachy. Graphical representation of the number of significant contributions to the chi-square statistics of the heterotachous sites per monophyletic group. Red, orange, and yellow stacks indicate evolutionary rates that are significantly faster than would be expected if there were no heterotachy, red indicating strong evidence for faster evolution and yellow indicating weak evidence. Green, blue, and purple stacks indicate evolutionary rates that are significantly slower than would be expected if there were no heterotachy, purple indicating strong evidence for slower evolution and green indicating weak evidence.

in this group (when $O_{ij} > E_{ij}$), in which case we labeled them "positive," or being unexpectedly low (when $O_{ij} < E_{ij}$), in which case we labeled them "negative." As seen in figure 2, Euglenida, Ciliophora, Platyhelminthes, and Annelida are the monophyletic groups that contribute the most to the presence of heterotachy, oftentimes due to sites evolving faster than expected.

Function–Structure Analysis

In Figure 3, the heterotachous sites were mapped on the secondary structure of the yeast *Saccharomyces cerevisiae*. We tested the distribution of the heterotachous sites over the different regions of the secondary structure such as stems, hairpin loops, internal loops, branching loops, and single-stranded regions (bulge loops and pseudoknots were not considered because they rarely occur). Percentages of heterotachy were 27.9% in stem regions, 30.5% in hairpin loops, 34.6% in the internal loops, 29.5% in branching loops, and 27.6% in the single-stranded regions. In line with previous studies (Philippe and Lopez 2001; Lopez et al. 2002), a chi-square test of homogeneity revealed no evidence for an uneven distribution of heterotachy between regions (*P* value = 0.87).

In the past, substitution rates of eukaryotic and bacterial SSU rRNAs have been superimposed on the secondary structure of *S. cerevisiae* (Wuyts et al. 2001) and both the secondary and tertiary structure of *Thermus thermophilus* (Van de Peer, Chapelle, and De Wachter 1996). Inspection of the substitution rates showed that structurally interacting sites in an RNA molecule evolve very similarly in virtually every case, apart for few exceptions. This is due to compensatory mutations and to the constraint of maintaining

the secondary structure of the rRNAs (Higgs 2000). A mutation on one side of a pair within a helical region disrupts the structure and is slightly deleterious, unless a second mutation of the other side of the pair restores the pairing ability (Savill et al. 2001).

To investigate whether sites at opposing sides of a stem evolve according to the same principle (heterotachy or homotachy), 220 bp (i.e., all the nucleotide pairs within the stem regions) in the secondary structure were evaluated. One hundred and sixty five (or 75%) evolved according to the same principle (both heterotachy or both homotachy), whereas 55 (or 25%) evolved according to a different principle. To assess the significance of this result, one must acknowledge that a significant proportion γ_0 of pairs may evolve according to the same principle just by chance (i.e., even when heterotachous sites are randomly distributed over the alignment). With π being the probability of observing a heterotachous site within the 220 pairs (i.e., the percentage of heterotachous sites among the 440 sites that make up the 220 bp), it can be shown that this proportion equals $\gamma_0 = \pi^2 + (1 - \pi)^2$. For our data, we estimated $\pi = 27\%$, suggesting that paired evolution happens by chance in 61% of all pairs. To assess whether the observed chance γ of paired evolution ($\hat{\gamma}=75\%$) is significantly elevated, we used the delta method to acknowledge that γ_0 is itself estimated (see Appendix A) and found a *P* value of 1.3×10^{-6}, suggesting highly significant evidence for paired evolution.

Degree of Support

Although in the absence of a molecular clock, the GTR evolutionary model used for fitting the data does
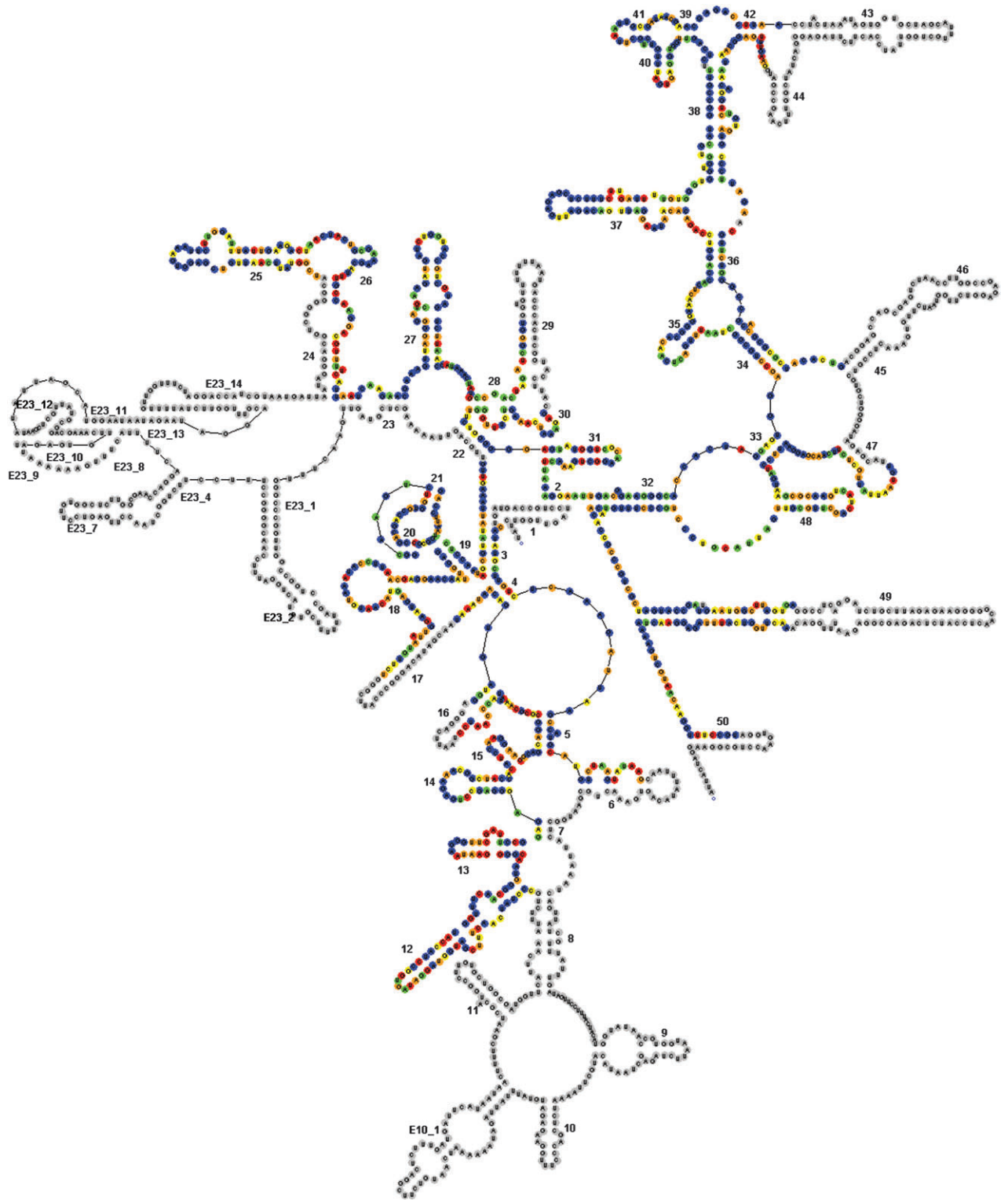
Fig. 3.—Distribution of heterotachy mapped on the SSU rRNA secondary structure of *S. cerevisiae*. The sites have been subdivided into 5 categories, expressing the degree of evidence in favor of heterotachy or homotachy. The heterotachous positions with the lowest *q* values (below 2.5%) are colored in red, the other heterotachous sites in orange. Remaining sites are in yellow (i.e., *q* value between 5% and 7.5%), green (i.e., *q* value between 7.5% and 10%), and blue (i.e., *q* value above 10%). Sites in gray were not present in our final alignment and could thus not be tested.

not prohibit sites to have different evolutionary rates in different lineages, the combination of heterotachous and homotachous sites in an alignment may generate inaccuracies in phylogenetic inference due to the heterogeneity of the sites (Moreira and Philippe 2000). One might therefore expect that the bootstrap support will tend to increase after removing heterotachous sites, creating a more homogeneous alignment and reducing model violations (Philippe and Germot 2000). Due to computational constraints, a random subset of 10 sequences of each monophyletic group was used to calculate the bootstrap supports with and without the removal of heterotachous sites. Comparing these bootstrap supports for the clustering of specific monophyletic groups revealed a surprising and unexpected systematic decrease for most monophyletic groups when heterotachous sites were removed, as is shown in table 2. A similar result was observed in a covarion study by Lockhart et al. (1998). Likewise, using simulation studies, Penny et al. (2001) have shown that the chance of recovering a correct tree topology increases when sites that are unchangeable for part of the time are present. Their results are therefore also suggestive of decreased bootstrap supports when removing "covarion" sites.

To acknowledge that this decrease in bootstrap support could be merely due to the decrease in number of available sites, regardless of the presence of heterotachy, control experiments are necessary to correctly determine the impact of the removal of heterotachous sites (Inagaki et al. 2004). To this end, we simulated $N = 100$ alignments from the original alignment, randomly removing an equal number (i.e., 283) of sites on each occasion. For each alignment, a bootstrap tree was constructed based on 5000 bootstrap replications, using the same procedure we used for obtaining the original trees. For each tree, the bootstrap value in each monophyletic group was determined. These bootstrap values were then ordered from smallest to largest: $M_{(1)} \leq \ldots \leq M_{(N)}$. To acknowledge simulation error due to using a finite number N of trees, we calculated an approximate 95% confidence interval (CI) for the 5% percentile as $[M_{(L)}, M_{(U)}]$ (Nettleton and Doerge 2000), where

$$L = \left\lceil 0.05N - \Phi^{-1}(0.975)(N(0.95)0.05)^{\frac{1}{2}} \right\rceil$$

and

$$U = \left\lceil 0.05N + \Phi^{-1}(0.975)(N(0.95)0.05)^{\frac{1}{2}} \right\rceil,$$

where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. We subsequently verified whether the number N of simulated alignments was sufficient to produce unambiguous results (Nettleton and Doerge 2000). Bootstrap values below the lower bound of this CI are unexpected and therefore suggestive of the decrease in bootstrap support not just being due to the random removal of sites.

In some cases (i.e., for some removals of 283 random sites), some monophyletic groups could not be recovered, and hence, their bootstrap value was not obtainable from standard software. To cope with this missing information, we constructed 2 CIs as in a worst-case/best-case analysis. In the former, we imputed 0 for the missing bootstrap support. In the latter, we chose the minimal bootstrap support encountered for the given monophyletic group across all available trees (a low bootstrap value is realistic because the group could not be reconstructed, that is, bootstrap support was very low). Table 2 shows that the bootstrap supports for the different groups, after removal of heterotachy, were systematically lower than $M_{(L)}$. We therefore conclude that the decrease in support after removal of heterotachy is not just due to the removal of random sites. It follows that, in addition to the removal of sites, another process must be responsible for the decrease in bootstrap support.

To investigate whether the decrease in bootstrap support is the result of a different variability at heterotachous sites than others, we subsequently constructed $N = 300$ alignments from an alignment from which random sites with the same variability were removed from the data set. To randomly select sites with the same variability, we first subdivided sites into different classes containing at least 10 different sites of similar variability, as measured by the number of substitutions at the considered site. Next, for each heterotachous site in each variability class, we randomly drew a (homotachous or heterotachous) site from the same variability class. From the results in table 2, we conclude that the decrease in bootstrap support may well be caused by the increased variability at detected heterotachous sites (which may itself be due to the increased power of our test at sites with high variability). This is seen because the bootstrap supports are systematically higher than $M_{(U)}$. Note that for some groups (e.g., Platyhelminthes), there was no decrease in bootstrap support when removing the heterotachous sites. Other groups (e.g., Mollusca and Apicomplexa) may require further testing because their initial bootstrap support (when constructing the tree using all sites) was considered insufficient (i.e., <70%).

Table 2
Significance Assessment of Decreased Bootstrap Values

Group	All pos.	HT rem.	Rand. WC	Rand. BC	Variable WC	Variable BC
Platyhelminthes	100	100		[99–100]		[98–100]
Euglenida	99	51		[65–85]		[31–50]
Cnidaria	99	78		[81–89]		[68–77]
Ascomycota	95	43		[51–63]	[0–42]	[19–42]
Ciliophora	92	36		[43–68]	[0–0]	[18–18]
Mollusca	56	38	[0–0]	[8–8]	[0–0]	[8–8]
Apicomplexa	56	21	[0–0]	[21–21]	[0–0]	[17–17]

NOTE.—Bootstrap values for specific monophyletic groups (first column) under different conditions. Second column, using all sites; third column, after removal of detected heterotachous sites; fourth column, after random removal of an equal number of sites (95% CIs for 5% percentile; WC, worst case; BC, best case); fifth column, after random removal of an equal number of sites with similar variability (95% CIs for 5% percentile).

Discussion

Since the introduction of models that aim to incorporate the covarion hypothesis (Tuffley and Steel 1998), there has been an increasing amount of research on both covarion and heterotachy processes. Recent work includes the comparison between the performance of maximum parsimony, maximum likelihood, and Bayesian Markov chain Monte Carlo using simulated data sets containing heterotachy (Kolaczkowski and Thornton 2004; Spencer et al. 2005). However, given the complexity of such evolutionary processes, limited research has been conducted to assess the impact on phylogenetic inference under biologically plausible models (Steel 2005). In this respect, we believe that our results, being based on real data, provide valuable insights into the actual patterns of heterotachy, as they have occurred through evolution in SSU rRNAs, and into their influence on the support of phylogenetic inference.

It is well known that the changes at the 2 sides of a stem region in RNA are correlated with each other due to the constraint of maintaining the secondary structure (Higgs 2000). Our method to detect heterotachous sites significantly confirms such a correlation, which is suggestive of its adequate performance. We expect this similarity to be even more pronounced in reality. This is because the maximum likelihood approach used for inferring the evolutionary rates treats sites within base pairs as independent. Although this approach is known for its robustness when violating interdependencies of sites, it remains to be investigated whether results would modify if rates were inferred using base-pair models such as a 16-state Markov model (Schöniger and von Haeseler 1994), a 7-state (Tillier and Collins 1998), or 6-state model (Tillier and Collins 1995). Note also that there may always be base pairs containing one heterotachous site on one side and one homotachous site on the other side of the stem. This may happen in a situation with 2 Watson–Crick pairings (A-U and G-C) and one non–Watson–Crick interaction (G-U) (Gutell et al. 1992), as in the model of Tillier and Collins (Tillier and Collins 1995). Indeed, it is possible that in a given monophyletic group, the transition from G-C to G-U has occurred several times but that the base pair mutates back to G-C instead of selecting the compensatory mutation to A-U. This could imply that sites at opposing sides of the stem show differences in variability, which could result (accumulated over different groups) in the detection of heterotachy instead of homotachy (or vice versa) at opposing sites of a stem region (fig. 3).

In our analysis of the secondary structure of the SSU rRNA, we found no immediate correlation of heterotachy and structure–function for SSU rRNA. Although it has been shown that modeling of heterotachy may provide higher likelihoods in the reconstruction of phylogenetic trees (Galtier 2001), the role of the heterotachy process in structural and functional research still remains unclear.

Future Work

Similar to other approaches for detecting heterotachy, our method detects heterotachy between different evolutionary lineages (Lopez et al. 2002; Kolaczkowski and Thornton 2004). However, the definition of heterotachy allows evolutionary rates to change more generally through time, that is, across the phylogeny. This means that a rate switch can occur anywhere in the phylogenetic tree and not only at the internal nodes between kingdoms or major phylogenetic clades. Our heterotachy test cannot be used to detect rate switches at a chosen branch of the tree. Adaptations that allow more flexibility will likely lead to increased power for the heterotachy test. Further, preliminary analyses have shown that the tree length of each monophyletic group is an adequate measure for the overall evolutionary rate (i.e., an adequate choice of λ_i) for our chi-square test. Nonetheless, other measures might prove useful for certain data sets and possibly lead to greater power for the heterotachy test. Regardless of how one measures the overall evolutionary rate λ_i, it will typically be based on estimated substitution numbers. It remains to be investigated how the resulting uncertainty about λ_i may impact our test results. As in other research studies concerning evolutionary rates, computational constraints make this currently prohibiting, however.

Recent studies have focused on modeling covarion and SSRV (i.e., heterotachy), but current evolutionary models for these processes are limiting. The covarion hypothesis is usually modeled by superimposing 2 stochastic processes: a 2-state Markov process that acts as a switch, turning sites "on" (variable) and "off" (invariable), and a standard substitution process for sites in "on" state, corresponding to an evolutionary model of choice (Tuffley and Steel 1998; Huelsenbeck 2002). Likewise, heterotachy has been modeled by superimposing a continuous rate-changing process onto the among-site rate variation nucleotide process (Galtier 2001). Both models were found to provide equally well or better fits to the data in most cases, as compared with nucleotide models that do not allow SSRV. However, these models assume site-independent evolution, an assumption that contradicts the covarion hypothesis as formulated by Fitch and Markowitz (1970). It thus remains to be investigated how one can model the typical behavior of covarions, that is, when a mutation gets fixed at a certain position, another position becomes variable.

Conclusion

In this study, we have proposed a statistical method to uncover heterotachy in an alignment involving a priori identified monophyletic groups. In a data set of 1289 aligned eukaryotic SSU rRNA sequences, we estimated that heterotachy is present at 66.3% of the sites. In addition, our method identified 29.2% of the sites to be heterotachous when controlling the FDR at 5%. No evidence was found that these sites were heterogeneously distributed along the SSU rRNA; that is, there is no evidence that the secondary structure directly affects the probability for a position to be heterotachous. We showed that sites at opposing sides in the stem regions evolve similarly, as expected for stem regions within RNA. We extensively investigated the effect of heterotachous sites on the support for certain branchings within trees reconstructed using evolutionary models without SSRV. We observed that the removal of heterotachous sites leads to decreased bootstrap supports and showed that

this may be explained by the increased variability at heterotachous sites.

Appendix A

Let Y_i equal 1 in the case of heterotachy at position i (i.e., position i has a q value below 5%) and 0 in the case of homotachy (i.e., position i has a q value above 5%). Let X_i equal 1 when there is similar evolution base pair i (i.e., if both paired sites have either a q value below 5% or they both have a q value above 5%) and 0 otherwise. Using Y and X, the following estimators for γ and π were constructed:

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{and} \quad \hat{\pi} = \frac{\sum_{i=1}^{n} Y_i}{2n}.$$

The test statistic that we used to assess whether $\gamma \neq \pi^2 + (1 - \pi)^2$ is defined as

$$\hat{\theta} = g(\hat{\gamma}, \hat{\pi}) = \hat{\gamma} - \hat{\pi}^2 - (1 - \hat{\pi})^2,$$

where $g(\gamma, \pi) = \gamma - \pi^2 - (1 - \pi)^2$ and $(\hat{\gamma}, \hat{\pi})$ has the following approximate distribution:

$$(\hat{\gamma}, \hat{\pi}) \sim N((\gamma, \pi), \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \frac{\gamma(1-\gamma)}{n} & \text{cov}(\hat{\gamma}, \hat{\pi}) \\ \text{cov}(\hat{\gamma}, \hat{\pi}) & \frac{\pi(1-\pi)}{n} \end{pmatrix}$$

and

$$\text{cov}(\hat{\gamma}, \hat{\pi}) = \frac{\sum_{i=1}^{n} \text{cov}(X_i, Y_i)}{(2n)^2} = \frac{\text{cov}(X_i, Y_i)}{2n}.$$

Using the delta method (see for instance, Wasserman 2004), we then find that approximately

$$\hat{\theta} = g(\hat{\gamma}, \hat{\pi}) \sim N(g(\gamma, \pi), g'(\gamma, \pi)^T \sum g'(\gamma, \pi))$$

under the null hypothesis that $\gamma = \gamma_0$.

For our data, we find that $\hat{\theta} = 0.14$ and that $\hat{\theta} \sim N(0; 9.47 \times 10^{-4})$ under the null hypothesis. This gives a P value of 1.3×10^{-6}, indicating significant evidence for paired evolution.

Appendix B

In this appendix, we provide the general algorithm for estimating q values from a list of P values, as it appeared in Storey and Tibshirani (2003)

1. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ be the ordered P values. This also denotes the ordering of the features in terms of their evidence against the null hypothesis.
2. For a range of λ, say $\lambda = 0, 0.01, 0.02, \ldots, 0.95$, calculate $\hat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1-\lambda)}$. When the P values are not uniformly distributed under the null hypothesis, one should instead use our approach from the section Assumption of Uniform P values.
3. Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ.
4. Let the estimate of π_0 be $\hat{\pi}_0 = \hat{f}(1)$.

5. Calculate $\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \hat{\pi}_0 p_{(m)}$.
6. For $i = m-1, m-2, \ldots, 1$, calculate the estimated q value for the ith most significant feature to be
$$\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \min\left(\frac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right).$$

Appendix C

Here, we provide an artificial example of how the substitutions can be redistributed over the different monophyletic groups and sites during a permutation. Table 3 contains substitutions for 4 sites (1 through 4) within 3 groups, as well as their totals.

Table 3
Original Substitutions

Site	1	2	3	4	Tree Lengths
Group 1	5	7	2	3	17
Group 2	2	4	0	4	10
Group 3	1	0	1	0	2
Total	8	11	3	7	29

To redistribute the substitutions, we proceed from left to right starting with 8 substitutions to redistribute over the 3 different groups. Because we wish to keep the tree length fixed, the first substitution has a chance of 17/29 = 59% to be assigned to group 1, 10/29 = 34% to be assigned to group 2, and 2/29 = 7% to be assigned to group 3. Generating a random number between 1 and 29 indicates to which group a first substitution should be assigned: if the number is between 1 and 17 the substitution is assigned to the first group, if it is between 18 and 27 the substitution is assigned to the second group, and if the number is 28 or 29 the substitution is assigned to the third group. After each assignment of a substitution to a group, the tree length of that group is decremented. For example, should the first substitution be assigned to the first group, the tree length of that group would be decremented to 16. Next, we proceed to the following substitution. This way, both tree lengths and the total amount of substitutions per site will remain fixed, resulting in a possible permutation as illustrated in Table 4 below.

Table 4
Permutation

Site	1	2	3	4	Tree Lengths
Group 1	8	5	1	3	17
Group 2	0	5	2	3	10
Group 3	0	1	0	1	2
Total	8	11	3	7	29

Literature Cited

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57:289–300.

Fitch WM. 1971. Rate of change of concomitantly variable codons. J Mol Evol 1:84–96.

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 4:579–93.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–73.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–74.

Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD. 1992. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res 20:5785–95.

Higgs PG. 2000. RNA secondary structure: physical and computational aspects. Q Rev Biophys 33:199–253.

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

Inagaki Y, Susko E, Fast NM, Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and archaebacteria in EF-1α phylogenies. Mol Biol Evol 21:1340–9.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–4.

Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol 23:40–5.

Lockhart P, Steel M. 2005. A tale of two processes. Syst Biol 54:948–51.

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 93:1930–4.

Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol 15:1183–8.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process in 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 in insects: site-specific rate variation is clustered and frequently detected. J Mol Evol 55:460–9.

Moreira D, Philippe H. 2000. Molecular phylogeny: pitfalls and progress. Int Microbiol 3:9–16.

Nettleton D, Doerge RW. 2000. Accounting for variability in the use of permutation testing to detect quantitative trait loci. Biometrics 56:52–8.

Olsen GJ. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol 52:825–37.

Penny D, McComish BJ, Charleston MA, Hendy MD. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol 53:711–23.

Philippe H, Germot A. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol Biol Evol 17:830–4.

Philippe H, Lopez P. 2001. On the conservation of protein sequences in evolution. Trends Biochem Sci 26:414–6.

Roff DA, Bentzen P. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol Biol Evol 6:539–45.

Savill NJ, Hoyle DC, Higgs PG. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics 157:399–411.

Schöniger M, von Haeseler A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol Phylogenet Evol 3:240–7.

Spencer M, Susko E, Roger AJ. 2005. Likelihood, parsimony and heterogeneous evolution. Mol Biol Evol 22:1161–4.

Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'? Trends Genet 21:307–9.

Steel M, Huson D, Lockhart PJ. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst Biol 49: 225–32.

Storey JD. 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31:2013–35.

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 100:9440–5.

Swofford DL. 2001. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sunderland, MA: Sinauer Associates.

Tillier ERM, Collins RA. 1995. Neighbour joining and maximum likelihood with rna sequences: addressing the interdependence of sites. Mol Biol Evol 12:7–15.

Tillier ERM, Collins RA. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics 148:1993–2002.

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

Van de Peer Y, Ben Ali A, Meyer A. 2000. Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. Gene 246:1–8.

Van de Peer Y, Chapelle S, DeWachter R. 1996. A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Res 24:3381–91.

Van de Peer Y, Rensing SA, Maier U-G, De Wachter R. 1996. Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc Natl Acad Sci USA 93:7732–6.

Wasserman L. 2004. All of statistics: a concise course in statistical inference. New York: Springer.

Wuyts J, Perrière G, Van de Peer Y. 2004. The European ribosomal RNA database. Nucleic Acids Res 32:D101–3.

Wuyts J, Van de Peer Y, De Wachter R. 2001. Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. Nucleic Acids Res 29:5017–28.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol Evol 11:367–70.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6. Available from: http://abacus.gene.ucl.ac.uk/software/paml.html.