Copyright © 1998 by the Genetics Society of America

High Apparent Rate of Simultaneous Compensatory Base-Pair Substitutions in Ribosomal RNA

Elisabeth R. M. Tillier and Richard A. Collins

Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A8

Manuscript received September 29, 1997
Accepted for publication December 17, 1997

ABSTRACT

We present a model for the evolution of paired bases in RNA sequences. The new model allows for the instantaneous rate of substitution of both members of a base pair in a compensatory substitution (*e.g.*, A-U→G-C) and expands our previous work by allowing for unpaired bases or noncanonical pairs. We implemented the model with distance and maximum likelihood methods to estimate the rates of simultaneous substitution of both bases, α_d, *vs.* rates of substitution of individual bases, α_s in rRNA. In the rapidly evolving D2 expansion segments of Drosophila large subunit rRNA, we estimate a low ratio of α_d/α_s, indicating that most compensatory substitutions involve a G-U intermediate. In contrast, we find a surprisingly high ratio of α_d/α_s in the core small subunit rRNA, indicating that the evolution of the slowly evolving rRNA sequences is modeled much more accurately if simultaneous substitution of both members of a base pair is allowed to occur approximately as often as substitution of individual bases. Using simulations, we have ruled out several potential sources of error in the estimation of α_d/α_s. We conclude that in the core rRNA sequences compensatory substitutions can be fixed so rapidly as to appear to be instantaneous.

ALTHOUGH ribosomal RNA genes have been extensively used for inferring the phylogeny of distantly related sequences (for reviews see Hillis and Dixon 1991 and Olsen and Woese 1993), several authors have pointed out that these sequences violate an assumption of the methods usually used to do the phylogenetic analysis: that the sites in the sequence evolve independently (Wheeler and Honeycutt 1988; Dixon and Hillis 1993; Tillier 1994; Tillier and Collins 1995; Schoniger and von Haeseler 1994; Muse 1995; Rzhetsky 1995). In rRNA, roughly half of the bases in the sequence are involved in intramolecular base pairs that form the characteristic secondary and tertiary structure required for function (Vawter and Brown 1993; Noller 1984), and the genes are under evolutionary constraint to maintain these pairings. Even though secondary structures are composed mostly of Watson-Crick base pairs, other base combinations are sometimes found: G-U pairs are especially common, and G-U is generally thought to be an evolutionarily stable intermediate in the substitution of a G-C pair with an A-U pair via two consecutive transitions (*e.g.*, Rousset *et al.* 1991).

To address the complexities imposed by RNA secondary structure, several authors have recently proposed different probability-based models for the evolution of paired bases in RNA sequences (Tillier 1994; Tillier and Collins 1995; Schoniger and von Haeseler 1994; Muse 1995; Rzhetsky 1995). These models vary in the number of possible states a base pair may have, and in the rates of substitution between these states; however, most of these models still assume that the substitution of each base occurs independently of its partner according to a Poisson distribution (pairing is favored in these models by having a higher rate of substitution from unpaired to paired than from paired to unpaired). Compensatory base-pair substitutions (*e.g.*, A-U→U-A) are still assumed to be the result of two independent events; their probabilities in a small amount of time are essentially nil and are therefore disregarded (Muse 1995; Rzhetsky 1995; Schoniger and von Haeseler 1994).

In contrast, we have developed a model for the evolution of double-stranded RNA sequences that allows for the complete evolutionary dependence by permitting the simultaneous substitution of both members of a base pair (Tillier 1994; Tillier and Collins 1995). Our model did not allow for base combinations other than Watson-Crick and G-U base pairs and is therefore not generally applicable for the analysis of actual RNA sequences. Here, we expand our previous model to allow for all possible base combinations. The expanded model permits us to estimate the rates of substitution of individual members of a base pair. More importantly, our model allows us to estimate the rate of simultaneous compensatory substitutions, and thereby test the validity of the commonly-made assumption that simultaneous substitutions do not occur.

Here, we use our expanded model to estimate the rates of substitution of base pairs and of individual bases, using a large data set of published sequences of small

Corresponding author: Richard A. Collins, Department of Molecular and Medical Genetics, Faculty of Medicine, University of Toronto, 1 King's College Circle, Toronto, Ontario, Canada M5S 1A8. E-mail: rick.collins@utoronto.ca

subunit (SSU) rRNA. We cannot estimate absolute values of the rate parameters irrespective of time because, in most cases, the true time of divergence is unknown, but we can estimate their relative rates. Of particular interest is the ratio of the rate of the simultaneous substitution of both members of the base pair to the rate of substitution to and from the G-U transitional intermediate. This quantity addresses the question of whether G-U is necessarily an evolutionarily stable intermediate in the substitution between G-C↔A-U pairs. We find that this ratio is substantially greater than zero; in fact, the evolution of rRNA sequences is modeled far more accurately if simultaneous substitution of both members of a base pair is allowed to occur approximately as often as substitution of individual bases.

THE MODEL

The description of the probability of substitution of any base combination to any other is mathematically complicated, involving a 16 × 16 matrix and potentially as many rate parameters as there are types of substitutions. Different authors have used different simplifying assumptions to limit the number of independent rate parameters to three or fewer (Muse 1995; Rzhetsky 1995; Schoniger and von Haeseler 1994). In a previous paper (Tillier and Collins 1995), a novel model of base-pair substitution was examined to determine whether the then-current methods of phylogenetic analysis, which assumed that bases evolve independently, are appropriate for analysis of RNA sequences such as rRNA, where this assumption does not hold. We developed a model of base substitution for use with the Neighbor Joining and Maximum Likelihood methods that takes into account the evolutionary dependence of the sites in the molecule. We previously chose to first limit the size of the matrix by only considering A-U, G-C and G-U base pairs, but this led to a very strict definition of a paired site, because any site containing any other base combination would not be considered within this model. An important implication is that there were no unpaired intermediates considered in the model to allow for single base substitutions, except for G-U. Therefore, the model assumed that other unpaired intermediates are eliminated so quickly that both members of a base pair are apparently substituted simultaneously, an event we call a double substitution. In fact, the model proposed an instantaneous rate of double substitution, α_d, from A-U to G-C, even when single base transitions at a rate α_s through a G-U intermediate were possible. Because all single transversion mutations would disturb base pairing, all transversions were considered double compensatory substitutions, occurring at an instantaneous rate β.

To more accurately describe the evolution of real rRNA sequences, the model previously presented (Tillier 1994; Tillier and Collins 1995) was expanded to allow for the occasional occurrence of base combinations other than Watson-Crick and G-U (Figure 1). Because the 16 × 16 matrix of all possible base combinations with three or four rates of substitution become mathematically cumbersome with so many parameters, the model was simplified by collapsing all the other base combinations not considered in the original model as one type called O-T (for "other") which has an equilibrium frequency OT or π_7. Substitutions to or from O-T base combinations are allowed at an equal rate γ. We call this model the Other RNA Model, or OTRNA. This model reduces to our previous model (Tillier 1994; Tillier and Collins 1995), which we will refer to as the RNA Model when $\pi_7 = 0$. A diagram of the OTRNA Model is given in Figure 1, and the transition probability matrix, derived as in Tillier (1994), is given in Figure 2.

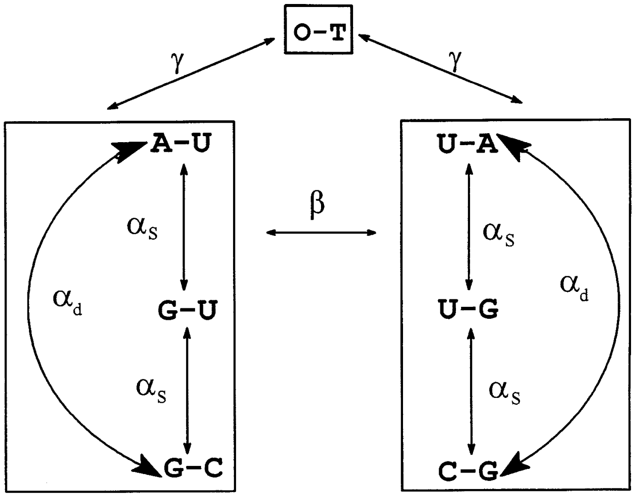

Figure 1.—The OTRNA Model. This is a schematic representation of an instantaneous substitution model. There are seven possible states for a site: A-U, G-U, G-C, U-A, U-G, C-G, O-T with frequencies π_1, π_2, π_3, π_4, π_5, π_6, and π_7, which are considered to be at equilibrium. We also define $\pi_1 + \pi_4 = \pi_8$, $\pi_2 + \pi_5 = \pi_9$, $\pi_3 + \pi_6 = \pi_{10}$. The base combinations can be considered in three groups each enclosed by a box. In the first two groups, single transition substitutions occur at a rate α_s to or from the G-U base pair and double transition substitutions occur at a rate α_d from A-U to G-C and vice versa. Double transversion substitutions occur between these two groups at rate β. All other substitutions occur to and from another combination of bases lumped together as O-T with a rate γ. The α_s, α_d, β, and γ are the number of substitutions per unit time.

ESTIMATING THE RATIO OF DOUBLE TO SINGLE SUBSTITUTION RATES

Distance estimate: The transition probability matrix allows us to estimate the parameters in the model multiplied by the time of divergence from the observed number of differences between any two sequences. The use of a probability model corrects for multiple substitu-

$$\boldsymbol{P}_{OTRNA}(t) =$$

$$\begin{array}{c|ccc}
 & AU & GU & GC \\
\hline
AU & \pi_1(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_2}{\pi_1+\pi_3}f_3)+\pi_3f_4 & \pi_2(1+f_1\frac{\pi_7}{1-\pi_7}+\pi_2f_2-f_3) & \pi_3(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_2}{\pi_1+\pi_3}f_3-f_4) \\
GU & \pi_1(1+f_1\frac{\pi_7}{1-\pi_7}+f_2-f_3) & \pi_2(1+f_1\frac{\pi_7}{1-\pi_7}+f_2)+(\pi_1+\pi_3)f_3 & \pi_3(1+f_1\frac{\pi_7}{1-\pi_7}+f_2-f_3) \\
GC & \pi_1(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_2}{\pi_1+\pi_3}f_3-f_4) & \pi_2(1+f_1\frac{\pi_7}{1-\pi_7}+f_2-f_3) & \pi_3(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_2}{\pi_1+\pi_3}f_3)+\pi_1f_4 \\
UA & \pi_1(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_2(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_3(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) \\
UG & \pi_1(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_2(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_3(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) \\
CG & \pi_1(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_2(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_3(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) \\
OT & \pi_1f_1 & \pi_2f_1 & \pi_3f_1
\end{array}$$

$$\begin{array}{cccc|}
UA & UG & CG & OT \\
\pi_4(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_5(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_6(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_7(1-f_1) \\
\pi_4(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_5(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_6(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_7(1-f_1) \\
\pi_4(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_5(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_6(1+f_1\frac{\pi_7}{1-\pi_7}-f_2) & \pi_7(1-f_1) \\
\pi_4(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_5}{\pi_4+\pi_6}f_3)+\pi_6f_4 & \pi_5(1+f_1\frac{\pi_7}{1-\pi_7}+f_2-f_3) & \pi_6(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_5}{\pi_4+\pi_6}f_3-f_4) & \pi_7(1-f_1) \\
\pi_4(1+f_1\frac{\pi_7}{1-\pi_7}+f_2-f_3) & \pi_5(1+f_1\frac{\pi_7}{1-\pi_7}+f_2)+(\pi_4+\pi_6)f_3 & \pi_6(1+f_1\frac{\pi_7}{1-\pi_7}+f_2-f_3) & \pi_7(1-f_1) \\
\pi_4(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_5}{\pi_4+\pi_6}f_3-f_4) & \pi_5(1+f_1\frac{\pi_7}{1-\pi_7}+f_2-f_3) & \pi_6(1+f_1\frac{\pi_7}{1-\pi_7}+f_2+\frac{\pi_5}{\pi_4+\pi_6}f_3)+\pi_4f_4 & \pi_7(1-f_1) \\
\pi_4f_1 & \pi_5f_1 & \pi_6f_1 & (1-\pi_7)f_1
\end{array}$$

$$\begin{aligned}
f_1 &= e^{-\gamma t} \\
f_2 &= \frac{1}{1-\pi_7}\, e^{-(\gamma\pi_7+(1-\pi_7)\beta)t} \\
f_3 &= \frac{2}{1-\pi_7}\, e^{-(\gamma\pi_7+½(1-\pi_7)(\beta+\alpha_s))t} \\
f_4 &= \frac{2}{\pi_8+\pi_{10}}\, e^{-(\gamma\pi_7+½(1-\pi_7)(\beta+\pi_9\alpha_s+(\pi_8+\pi_{10})\alpha_d)t}
\end{aligned}$$

Figure 2.—Transition probability matrix for the OTRNA Model.

tions at a given site over time (Jukes and Cantor 1969). For the OTRNA Model, these quantities are given by

$$\gamma t = -\frac{1}{2}\ln\left(1 - \frac{U}{2\pi_7(1-\pi_7)}\right)$$

$$\beta t = -\frac{1}{2(1-\pi_7)}\ln\left(1 - \frac{2Q}{1-\pi_7} - \frac{U}{2(1-\pi_7)}\right) - \frac{\pi_7\gamma t}{1-\pi_7}$$

$$\alpha_s t = -\frac{1}{1-\pi_7}\ln\left(1 - \frac{Q}{1-\pi_7} - \frac{S(1-\pi_7)}{2\pi_9(\pi_8+\pi_{10})} - \frac{U}{2(1-\pi_7)}\right) - \beta t - \frac{2\pi_7\gamma t}{1-\pi_7}$$

$$\alpha_d t = -\frac{1}{(\pi_8+\pi_{10})(1-\pi_7)}\ln\left(1 - \frac{Q}{(1-\pi_7)} - \frac{S}{2(\pi_8+\pi_{10})} - \frac{D(\pi_8+\pi_{10})}{2\pi_8\pi_{10})} - \frac{U}{2(1-\pi_7)}\right) - \frac{\pi_9\alpha_s t}{(\pi_8+\pi_{10})} - \frac{\beta t}{(\pi_8+\pi_{10})} - \frac{2\pi_7\gamma t}{(\pi_8+\pi_{10})(1-\pi_7)}, \qquad (1)$$

where U, Q, S, and D are, respectively, the proportions of paired to unpaired, double transversions, single transitions, and double transition substitutions. The base-pair frequencies are given by the πs as defined in the legend to Figure 1.

From Equations 1, we can obtain the new distance measure with this model because

$$K_d t = 4(1-\pi_7)\pi_7\gamma t + (1-\pi_7)^2\beta t + 2\pi_9(\pi_8+\pi_{10})\alpha_s t + 2\pi_8\pi_{10}\alpha_d t. \qquad (2)$$

The approximate variances of these rate estimates due to sampling, as well as the approximate variance for the ratio of any two of these estimates, were also obtained. For example, the variance for α_d/α_s is given by

$$\sigma^2\left(\frac{\alpha_d t}{\alpha_s t}\right) = \frac{1}{2K}\,(a^2Q + b^2S + c^2D + d^2U - (aQ + bS + cD + dU)^2)$$

$$a = \frac{1}{\alpha_s t}\frac{\partial(\alpha_d t)}{\partial Q} - \frac{\alpha_d t}{(\alpha_s t)^2}\frac{\partial\alpha_s t}{\partial Q}$$

$$b = \frac{1}{\alpha_s t}\frac{\partial(\alpha_d t)}{\partial S} - \frac{\alpha_d t}{(\alpha_s t)^2}\frac{\partial\alpha_s t}{\partial S}$$

$$c = \frac{1}{\alpha_s t}\frac{\partial(\alpha_d t)}{\partial D} - \frac{\alpha_d t}{(\alpha_s t)^2}\frac{\partial\alpha_s t}{\partial D}$$

$$d = \frac{1}{\alpha_s t}\frac{\partial(\alpha_d t)}{\partial U} - \frac{\alpha_d t}{(\alpha_s t)^2}\frac{\partial\alpha_s t}{\partial U}, \qquad (3)$$

where the derivatives in the equation are obtained using Equations 1, and K is the number of bases.

To estimate α_d/α_s, we considered a large data set consisting of 473 bacterial and archaebacterial SSU rRNA [from the Ribosomal Database Project (RDP); Maidak *et al.* 1996]. The sequences had been aligned with the reference secondary structure of *E. coli* (Gutell 1994). Only the 461 base pairs that are found in this reference secondary structure are considered. All pair-wise comparisons were done to determine α_d/α_s. In order to limit the number of comparisons, and because there are wide variations in the base-pair composition of the sequences, suggesting that the base composition of the data set is not at equilibrium, we only considered the pair-wise comparisons of sequences that had similar base-pair composition (*i.e.*, where the base-pair frequencies of A-U, G-U and G-C did not differ by more than 5% between the two sequences in the comparison). The results from the 1044 comparisons are shown in Table 1 and Figure 3.

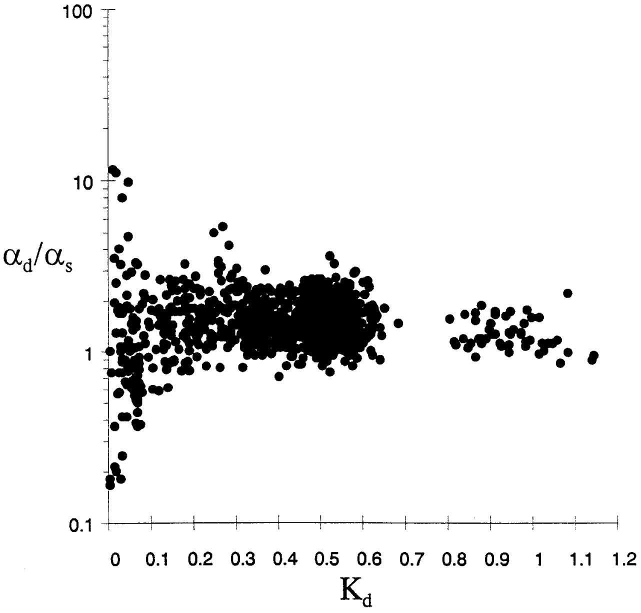

Figure 3.—Estimation of α_d/α_s from SSU rRNA data. The α_d/α_s estimate is plotted against the estimated distance (K_d) between any two sequences for all the pair-wise comparisons.

The estimated α_d/α_s ratio (1.45 ± 0.77 standard deviation, SD) is substantially greater than zero, and even larger than one. The sample SD is given but it is an underestimate of the true SD and cannot be used to test the significance of that mean. This is because the 1044 estimates are not independent, because the same sequence can be used more than once for different comparisons and because the phylogeny existing between these sequences is disregarded. The variance in the estimate is extremely high at low distances (<0.05) due to the lack of a sufficient number of substitutions to provide an accurate estimate. If the assumption of the OTRNA Model is met by these data, our finding that the rate of apparent simultaneous substitutions of base pairs is greater than the rate of single transition substitutions suggests that a large proportion of double transitions do not go through a G-U intermediate or

TABLE 1

Estimates of α_d/α_s with the OTRNA Model

	No. sequences	Distance estimate	Likelihood estimate
Eubacteria and archaebacteria	473	1.45 (0.77)	ND
Enteric bacteria	61	1.53 (0.80)	ND
	10	2.34 (0.58)	2.5
Ascomycetes (eukaryotes)	65	1.24 (0.64)	ND
	10	1.22 (0.37)	0.9

The table summarizes the estimates of α_d/α_s found with the different data sets using the pair-wise distance method and the likelihood method as described in the text (ND indicates the corresponding analysis was not done). The numbers in parentheses give the standard deviation for the pair-wise distance estimate.

that the G-U intermediate is lost before reaching fixation.

The other interesting rate that can now be estimated with the new model is the instantaneous rate of substitution from unpaired to paired bases (and vice versa) relative to the rate of double transversions. This ratio, β/γ, is estimated to be about 1.15 (SD 0.78). The variance in this estimate is high, and the quantity is difficult to estimate due to the rarity of transversions. Another difficulty with this quantity is that γ is a mixed rate to and from the various O-T base pairs, with most due to transversions and one due to transitions (by the C-A pair).

Possible sources of error in the distance estimate of α_d/α_s: *(1) Inadequate data set:* The data set used was large (473 sequences) and quite varied, consisting of sequences from the Eubacteria and Archaebacteria and covering the whole possible range of genetic distances. Because the variance in the estimate of α_d/α_s increases substantially with lower distances, and to determine whether there was any difference in the estimate of α_d/α_s from Eukaryotic sequences, we examined two more data sets. The first consisted of 61 Eubacterial sequences, all enteric bacteria aligned to the *E.coli* sequence. The second set consisted of 65 Eukaryotic sequences, all fungal ascomycetes, aligned to the *S. cerevisiae* sequence and reference secondary structure (Gutell 1994). All sequences were obtained from the RDP database (Maidak 1996). The range of K_d values for both sequence sets did not exceed 0.3. The estimated α_d/α_s for the bacterial data set was 1.53 (SD 0.8) and 1.24 (SD 0.64) for the ascomycetes (Table 1). Again, the variance in the estimate is high at the lower distances (not shown). These estimates do not vary substantially from the one previously found for the larger data set of 473 sequences (1.45, SD 0.77), increasing our confidence that α_d/α_s is greater than zero across all kingdoms.

(2) Lack of consideration of the phylogeny in the distance estimate of α_d/α_s: Because the estimation procedure used here does not take into consideration the phylogeny, it was necessary to determine whether the approach was valid. This question was addressed by computer simulations. 200 simulations were performed using the RNA Model as in the method described previously (Tillier 1994) following a tree leading to 82 sequences, as shown in Figure 4. The expected amount of divergence between the two most distantly related sequences was set to the very high value of $K_d = 1.8$, which is approximately the greatest value observed between an archaebacterium and the most distantly related eubacterium. Such a large value of K_d was used in order to magnify the effect of multiple substitutions so as to make it more easily detectable. Several lengths of sequences (numbers of sites) were used with values for the ratio α_d/α_s of 0 (the value expected if $\alpha_d = 0$) or 1, which is closer to the value that was approximated from the data. The value of α_d/β was set to two for these simulations because that is approximately the estimate for this quantity from the real data. The value of α_d/α_s was then estimated from the sequences by considering all pair-wise comparisons as was done with the real data. The results in Table 2 show that the estimate of α_d/α_s with this method is indeed a very accurate estimate of the value used to generate the data, even with a small number of sites. Even in the worst-case scenarios of very large distances and few sites, the simulations gave good estimates in situations which would give a high apparent number of double transitions that could possibly lead to an overestimation of α_d/α_s, suggesting that the pair-wise comparison approach, although crude, is nevertheless a valid one.

(3) Unknown mode of evolution: The real mode of evolution for rRNA sequences is of course unknown, and the OTRNA Model of evolution may be not be an accurate model of real evolution. Others (Schoniger and von Haeseler 1994; Muse 1995; Rzhetsky 1995) have proposed alternate models for the evolution of RNA sequences that do not allow for an instantaneous rate of double substitutions. If sequences actually evolved in ways that more resembled one of these alternate models, or any other way such that an instantaneous rate of double substitutions was not allowed, would our estimation procedure using the RNA Model still nevertheless estimate a large α_d/α_s? We used the Muse (1995) model for other simulations using the same tree (Figure 4), and again very high distances between the sequences.

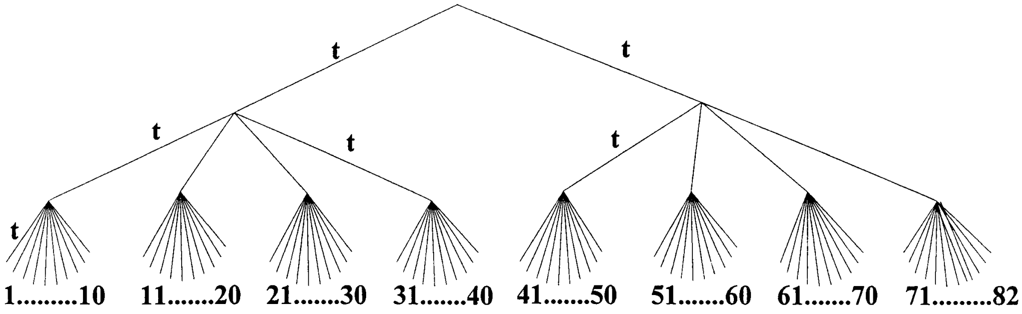

Figure 4.—82-species tree used in simulation. The lengths of the internodal branches in the tree are all equal to t, which was set in the simulation such that $K_d = 1.8$ between sequences 1 and 82.

The transition probability matrix was derived analytically by Rzhetsky (1995), for a more general model than that of Muse (1995), and was used for the simulation procedure. The model considers all 16 dinucleotide pairs but only has one parameter, λ, such that the rate of substitution from Watson-Crick pairs to all others is proportional to $1/\lambda$ and the reverse substitution rate proportional to λ. The rate of simultaneous substitution of both members of a dinucleotide pair is zero. We used the simplest version of the model where G-U is considered unpaired, and all base frequencies are equal. This model will yield a larger proportion of Watson-Crick pairs with increasing values of λ. We set $\lambda = 3$ for our simulations, which leads to expected equilibrium frequencies for A-U and G-C base pairs (π_8 and π_{10}) equal to 0.375, for G-U (π_9) equal to 0.0417 and all others (π_7) equal to 0.208. Any higher value for λ leads to a frequency of G-U too low to allow an accurate estimate of α_s (not shown). The sequences generated were analyzed with the OTRNA Model to estimate α_d/α_s. The result from 200 simulations is shown in Table 2, where the estimate of α_d/α_s is found to be essentially zero (−0.044 SD 0.017). The estimate is actually slightly negative because sequences evolved under the Muse model show fewer double changes that would be expected to be observed had sequences evolved under our model (even with $\alpha_d = 0$). The simulations therefore show that our method is robust, as they do not overestimate the value of α_d/α_s, even when the sequences have evolved following a very different model that does not allow for instantaneous double substitutions.

TABLE 2

Estimates of α_d/α_s from simulations

No.	RNA Model		Muse model
sites	$\alpha_d/\alpha_s = 0$	$\alpha_d/\alpha_s = 1$	$\lambda = 3$
25	0.080 (0.07)	1.090 (0.35)	ND
50	0.032	0.990	ND
75	0.013	0.987	ND
100	0.010	1.058	ND
200	−0.001 (0.02)	1.016 (0.19)	−0.044 (0.02)

Results of a simulation study investigating the accuracy of the distance method to estimate α_d/α_s for increasing length of sequence when the RNA Model and the model of Muse (1995) were used to simulate the evolution of sequences following the tree shown in Figure 4. The distance, K_d between the two most distant sequences was ~1.8. The values of α_d/α_s fixed for the simulations (0 and 1) with the RNA Model are indicated by the column headings. The entries in parentheses indicate the standard deviation of the estimated α_d/α_s for 200 replicates. ND, not done.

(4) Variable selection against G-U base pairs: Although the method seems to accurately take into account multiple substitutions at the sites, an uneven amount of selection against G-U base pairs at different sites in the sequence could create a systematic bias in the estimate of α_d/α_s. If some sites allow G-Us and others do not at all, then there are two types of sites in the molecule; sites at which G-Us are allowed and where transitions proceed mostly through G-U intermediates (type I sites), and sites where all substitutions are double (type II sites). We chose to investigate the degree of error in the estimate resulting from variable selection against G-U pairs using only the RNA Model because the effect would be more severe when O-T base pairs (particularly A-C, the other possible transitional intermediate) are not allowed. At type II sites, this model is equivalent to Hasegawa's two-parameter model (Hasegawa *et al.* 1985), where the rate of transition is α_d, the rate of transversion is β, and where we can have A-U, G-C, U-A or C-G rather than A, G, U or C at each site. Treating type I and type II sites separately allows for more accurate estimates of the parameters in the models if the number in each type of site is large enough.

It was possible to analytically determine the expected error in the estimate of α_d/α_s. This was performed by finding the expression for α_d/α_s when type I and type II sites are considered separately, and comparing it to the expression for the α_d/α_s obtained when the sites are all considered together. The relative error (the difference between the overestimate and the true value, divided by the latter) is plotted in Figure 5 against an increasing number of Type II sites and an increasing expected distance K_d between the sequences. The error and its standard deviation were also obtained by simulation. The graph shows that the relative error in the estimate of α_d/α_s is always less than 10%.

Maximum likelihood estimate: An independent way of estimating α_d/α_s is with Maximum Likelihood. The

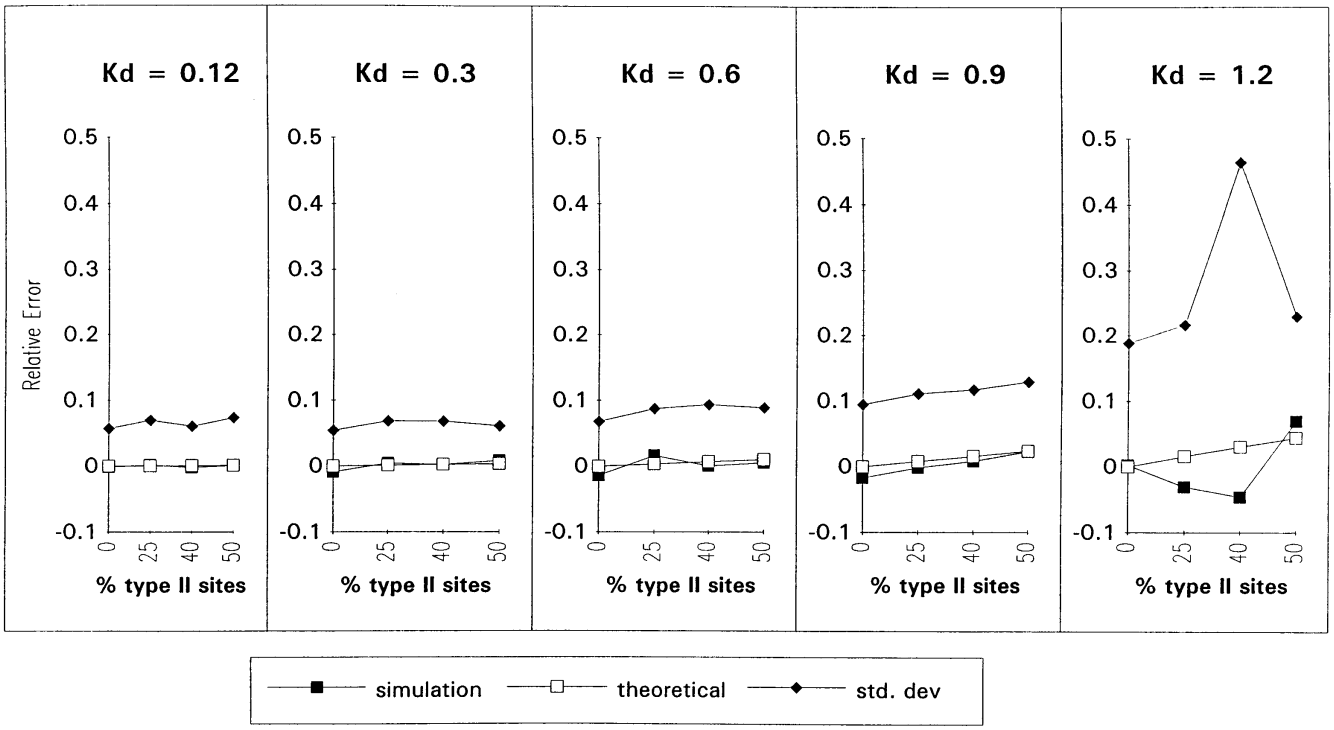

Figure 5.—Relative error in the estimate of the α_d/α_s ratio. This graph shows the systematic overestimation of the α_d/α_s ratio due to variable selection against G-U base pairs at different sites in the sequence. Each uninterrupted line in the graph corresponds to the error in the estimate obtained analytically (labeled theoretical in the legend) and with simulations or the standard deviation obtained by simulations, against an increasing proportion of type II sites (where G-U base pairs are not tolerated) in the 300 base-pair sequence (*i.e.*, each point corresponds to 0, 75, 120 or 150 type II sites in the 300 base-pair sequence, while the remainder are type I sites). There is one set of these graphs for an increasing amount of expected divergence between the two sequences.

likelihood ratio test, in which we compare a general model (either the RNA Model or the OTRNA Model) to a restricted model constructed by setting the α_d/α_s ratio to a constant. Interestingly, the restricted model, where α_d/α_s is fixed, does not have the mathematical properties of the more general model as described in Tillier (1994) that would allow us to use the EM maximization algorithm. A combination of Newton's method and the steepest descent method was therefore used for the maximization procedure.

Ten eubacterial sequences (*Escherichia coli, Buchera aphidicola, Citrobacter freundii, Erwinia herbicola, Serratia marcescens, Hafnia alvei, Rahnella aquatilis, Yersinia enterocolitica, Proteus vulgaris,* and *Plesiomonas shigelloides strain M51*; all enteric bacteria), chosen to be similar to the *E. coli* reference sequence to minimize changes in structure, were analyzed with the Maximum Likelihood approach. Another set of sequences, all fungal ascomycetes (*Saccharomyces cerevisiae, Blastomyces dermatitidis, Coccidioides immitis, Aspergillus fumigatus, Aureubasidium pullulans, Podospora anserina, Neurospora crassa, Colletotrichum gloesporioides, Torulaspora delbrueckii,* and *Schizosaccharomyces pombe*; aligned with the reference secondary structure of *S. cereisiae*), was also analyzed to determine whether the α_d/α_s estimate would also be large in the eukaryotic kingdom. The phylogenies and secondary structures for both sets were assumed to be those in the RDP database. The results of the likelihood ratio tests between the restricted models and the general model are plotted in Figure 6 for increasing values of the (fixed) α_d/α_s. Also plotted in Figure 6 are the results of a similar analysis performed using the OTRNA Model on two data sets obtained by simulation as in Tillier and Collins (1995). For the simulation analysis, sequences of similar length and base-pair frequencies to the eubacterial data set were generated under the same tree topology and with similar branch lengths as the real data tree. For the simulation, the OTRNA Model was restricted by setting α_d/α_s equal to either 0 or 2. The 95% confidence level is also shown on the graph, indicating the level below which the restricted model is not statistically different from the general model in explaining the data (*i.e.*, the values when the fixed α_d/α_s yields a maximum likelihood that is not significantly different when α_d/α_s is allowed to vary).

The curves for the analyses of both the bacterial and eukaryotic sequence data clearly show that $\alpha_d/\alpha_s > 0$ because fixing it at zero gives a significantly worse likelihood in all cases. This was observed no matter whether O-T base pairs are considered or not. The fixed α_d/α_s

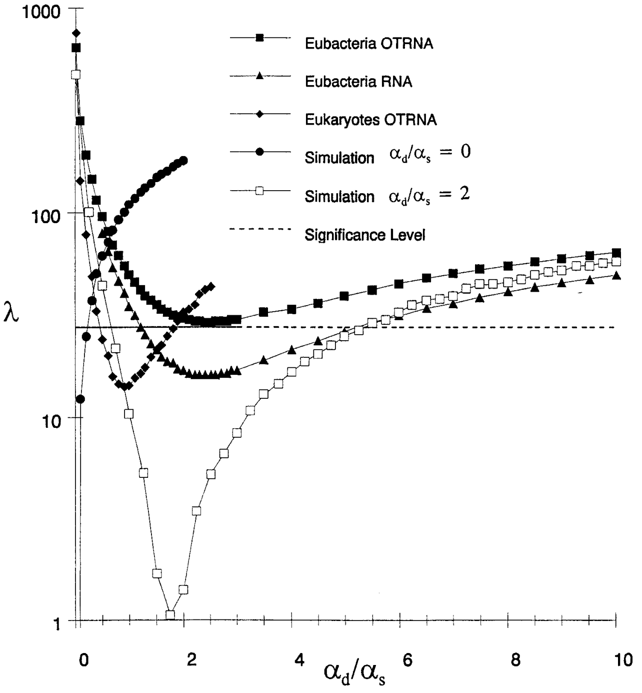

Figure 6.—Likelihood Ratio Tests. The graph shows the likelihood ratio statistic λ with increasing values of fixed α_d/α_s with several data sets of 10 sequences each.

that yielded the minimum value for the likelihood ratio with the eubacterial data set was at 2.3 using the RNA Model and 2.5 using the OTRNA Model. This is very close to the estimate of 2.3 obtained with the pair-wise distance method described above on this same data set modeled using the OTRNA Model. For the eukaryotic sequences, the minimum value of the likelihood ratio is found to occur at an α_d/α_s value of 0.9, also close to the value of 1.22 obtained using the pair-wise distance method. It is remarkable that the minima found for these curves are at or below the 95% confidence level, which implies that the α_d/α_s estimates are quite uniform from branch to branch, such that a uniformly imposed value is statistically justifiable.

The simulation curves in Figure 6 show that the method is valid, as the minima for these curves are very close to the value of α_d/α_s used in generating the data. Particularly telling is the extremely rapid rise in the likelihood ratio with increasing value of the estimated α_d/α_s when this ratio was actually set at zero in the simulation. This shows the likelihood method's ability to accurately estimate α_d/α_s when this quantity is zero. The minima in the curves for the simulated data sets fall well below the 95% confidence level as would be expected with simulated data being not as variable as real sequences, because in this case the only source of variation is due to sampling.

Interestingly, the value of α_d/α_s is estimated to be slightly higher when the OTRNA Model is used instead of the RNA Model, a phenomenon that is also observed when the distance method is used (data not shown). We would have expected the reverse result, as the OTRNA Model allows for the other transitional intermediate, C-A, whereas the RNA Model does not. The phenomenon may simply be because, when O-T base pairs are included, the relative frequency of G-U pairs is lowered and thus serves to increase the estimate of α_d.

Degree of independence and the effective number of sites: From the value of the rate parameters in the model we can obtain an estimate of the effective number of sites K_e, *i.e.*, the length of an equivalent sequence to the one considered if all sites were evolving independently.

$$K_e = k_i K = \frac{1}{2}\left(1 + \frac{\text{probability of a single base substitution}}{\text{probability of a base pair substitution}}\right)K. \qquad (4)$$

For the OTRNA Model, we have

$$K_e = \frac{1}{2}\left(1 + \frac{2\pi_9(\pi_8 + \pi_{10})\alpha_s + 4\pi_7(1 - \pi_7)\gamma}{2\pi_9(\pi_8 + \pi_{10})\alpha_s + 4\pi_7(1 - \pi_7)\gamma + 2\pi_8\pi_{10}\alpha_d + (1 - \pi_7)^2\beta}\right)K, \qquad (5)$$

where K is the number of bases. The multiplier of K, k_i, which we will refer to as "the independence factor" reflects the degree to which the individual bases in a base pair are evolving independently. The value of k_i can range from 1, indicating complete independence, to 0.5 indicating complete covariation. We can obtain an estimate of the parameters as described above (Equation 1) from RNA sequences. With the 473 sequence rRNA data set we obtain $k_i = 0.65$ (SD 0.04).

DISCUSSION

The models we have proposed allow for instantaneous rates of simultaneous substitution of both members of a base pair. It is often assumed that such rates, describing the probability of two concurrent unlikely events, are small enough to be negligible (Schoniger and von Haeseler 1994; Muse 1995; Rzhetsky 1995). However, analyzing rRNA sequences with our models, we found estimates of the apparent instantaneous double substitution rates relative to the rates of single substitution to be far from negligible. In fact, we found that, on average, the rate of double substitution was at least as large as the rate of single substitution.

We investigated several potential sources of error in estimating these rates, including the possibility that single base substitutions, over long periods of time, along with a low frequency of G-U base pairs, could lead to an overestimate of the rate of double transitions, particularly when the phylogenetic relationship between the sequences is disregarded. From simulation results (Figure 6 and Table 2), using sequences evolved under our model and that of Muse (1995) where instantaneous double substitutions were not allowed, it is clear that the methods are accurate in their estimates, even under the unfavorable conditions we described. Additionally, both likelihood and distance methods of obtaining the estimate of α_d/α_s with several sets of actual sequence data gave very similar results: α_d/α_s is much greater than zero.

The OTRNA Model, although more general than the previously proposed RNA Model (Tillier 1994; Tillier and Collins 1995), is still a very simplified description for the evolution of double-stranded RNA sequences. It does allow for all base combinations that could be intermediates between compensatory Watson-Crick substitutions and is therefore more useful for the estimation of the ratio of double *vs.* single substitution rates. One drawback of the OTRNA Model is that O-T to O-T base-pair substitutions are not considered as changes; however, a site at which O-T to O-T substitutions are frequent would probably be better described with a single-stranded model.

Another simplification of our model is the assumption that substitutions from paired sites to unpaired sites (or to G-U) and vice versa are considered to have the same rate (either α_s or γ) in either direction. In contrast, the models of Muse (1995) and Rzhetsky (1995) include the realistic assumption that losing a pairing would occur at a slower rate than gaining one back if the unpaired bases are selectively detrimental. However, their models make the assumption that the simultaneous substitution of both members of a base pair does not occur, which is not necessarily valid according to our analysis.

The present study and those of Muse (1995), Rzhetsky (1995), and Schoniger and von Haeseler (1994) are attempts to deal with the problem of base pairing in RNA sequences that go much further than the first analysis of Wheeler and Honeycutt (1988), who concluded that the pairings should be weighed by half or disregarded altogether. Dixon and Hillis (1993) gave a more sophisticated method for weighing the double-stranded sites that did not use a probability model. In this paper, we have shown how our probability model can also give an estimate of the degree to which the substitution of one member of a base pair is independent of its partner (Equation 4). This independence factor (k_i) could be thought of as a weighing factor for the double-stranded sites. We would not recommend using k_i as a weighing factor in a phylogenetic analysis, but it can be used to evaluate the degree of constraint in the RNA sequences due to the need to maintain pairing and thus the structure of the RNA. Multiplying the weighing factor by the number of bases in the sequence gives an estimate of the equivalent number of independently evolving sites in the molecule. In a previous paper (Tillier and Collins 1995), we showed that the confidence in the trees obtained in phylogenetic analyses is reduced when a double-stranded model is used, and this is largely due to the reduced number of effective sites.

The models proposed here for the double-stranded regions could be used in conjunction with other models for the single-stranded regions of the RNA in distance and maximum likelihood analyses for phylogenetic purposes. Rzhetsky (1995) and Muse (1995) combined models in this manner, and also included variation in rates at different sites along the molecule.

Rousset *et al.* (1991) analyzed the Drosophila large subunit rRNA D1 and D2 expansion segments sequences and concluded that A-U to G-C substitutions almost always go through a G-U intermediate. We performed a distance analysis using the OTRNA Model on the D2 segments (obtained from GenBank) and found that α_d/α_s is indeed low (0.17 SD 1.8), which agrees with the interpretation that the majority of A-U to G-C substitutions in these sequences involves a G-U intermediate. On the other hand, our analysis on the core of SSU rRNA sequences revealed a high α_d/α_s (approximately ≥ 1), suggesting that the majority of compensatory substitutions between Watson-Crick base-pairs does not involve a stable G-U intermediate. Our results provide direct support for the suggestion by Rousset *et al.* that conclusions based on rapidly evolving expansion segments should not be extrapolated to the slowly evolving core segments of rRNA. Although the secondary structure of the expansion segments appears to be conserved in Drosophila species, noncompensatory substitutions may be more tolerated by natural selection and therefore more evolutionarily stable than in the core rRNA. The degree of selective constraint on secondary structure will determine the evolutionary stability of noncompensatory, single substitutions. Indeed, we find that for the Drosophila D2 expansion segments, $k_i = 0.86$, indicating that the members of the base-paired sites in these sequences are on average evolving more independently than in core SSU rRNA sequences (where $k_i = 0.65$). Rousset *et al.* conclude from their study that a model of compensatory substitution that allows G-U intermediates is correct, but we find that such a model is too restrictive if it does not also allow for double substitutions that do not require any stable intermediate.

The rates estimated in this paper are substitution rates and therefore do not yield information on which, if any, of the many evolutionary forces (mutation, selection or drift) is dominant in bringing about the rapid rates of compensatory base-pair substitutions that are observed. High compensatory substitution rates could possibly be due to an increased mutation rate due to the palindromic nature of the stems in the RNA and, thus, in the coding DNA sequence, and to the phenomenon of templated mutations (see Golding 1987), although it is difficult to envision such a process in the case where paired sites can be separated by several hundred nucleotides. Alternatively, work by Kimura (1985) and more recently by Stephan (1996), using population genetic models, have shown that compensatory mutations can quickly become fixed in a finite population when recombination rates are low, as is the case within genes. A low frequency of deleterious or slightly deleterious intermediates may be present long enough for compensatory mutations to appear without the intermediates themselves ever being fixed. A high rate of substitution on a microevolutionary scale would appear instantaneous on a macroevolutionary scale.

We thank two anonymous reviewers for their constructive comments. Parts of this work were supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) postgraduate scholarship to E.T. Support was also provided by an NSERC research grant to R.C. R.C. is a fellow of the Canadian Institute for Advanced Reserach Program in Evolutionary Biology.

LITERATURE CITED

Dixon, M. T., and D. M. Hillis, 1993 Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. **10:** 256–267.

Golding, G. B., 1987 Nonrandom patterns of mutation are reflected in evolutionary divergence and may cause some of the unusual patterns observed in sequences, pp. 151–172 in *Genetic Constraints on Adaptive Evolution*, edited by V. Loeschcke. Springer-Verlag, Berlin.

Gutell, R. R., 1994 Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. Nucleic Acids Res. **22:** 3502–3507.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Hillis, D. M., and M. T. Dixon, 1991 Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66:** 411–453.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132, in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kimura, M., 1985 The role of compensatory neutral mutations in molecular evolution. J. Genet. **64:** 7–9.

Maidak, B. L., G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey *et al.*, 1996 The Ribosomal Database Project. Nucleic Acids Res. **24:** 82–85.

Muse, S. V., 1995 Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139:** 1429–1439.

Noller, H. F., 1984 Structure of ribosomal RNA. Annu. Rev. Biochem. **53:** 119–162.

Olsen, G. J., and C. R. Woese, 1993 Ribosomal RNA: a key to phylogeny. FASEB J. **7:** 113–123.

Rousset, F., M. Pélandakis, and M. Solignac, 1991 Evolution of compensatory substitutions through a G-U intermediate state in Drosophila rRNA. Proc. Natl. Acad. Sci. USA **88:** 10032–10036.

Rzhetsky, A., 1995 Estimating substitution rates in ribosomal RNA genes. Genetics **141:** 771–783.

Schoniger, M., and A. von Haeseler, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phyl. Evol. **3:** 240–247.

Stephan, W., 1996 The rate of compensatory evolution. Genetics **144:** 419–426.

Tillier, E. R. M., 1994 Maximum Likelihood with multi-parameter models of substitution. J. Mol. Evol. **39:** 409–417.

Tillier, E. R. M., and R. A. Collins, 1995 Neighbor Joining and Maximum Likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. **12:** 7–15.

Vawter, L., and W. M. Brown, 1993 Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics **134:** 597–608.

Wheeler, W. C., and R. L. Honeycutt, 1988 Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. Mol. Biol. Evol. **5:** 90–96.

Communicating editor: G. Brian Golding