Neighbor Joining and Maximum Likelihood with RNA Sequences: Addressing the Interdependence of Sites

Elisabeth R. M. Tillier and Richard A. Collins*

*Department of Botany and Department of Molecular and Medical Genetics, University of Toronto and Canadian Institute for Advanced Research Program in Evolutionary Biology

Intrastrand base pairings give ribosomal and other RNA molecules characteristic structures that are important for their function. In order to maintain these structures, a substitution at one paired site may have to be compensated for by an appropriate substitution at the complementary site. Thus paired sites do not evolve independently of one another. Most current methods for inferring phylogeny from molecular sequences assume that the sites are independent and will therefore give statistically unreliable and possibly erroneous results when used on structured RNA sequences. We analyze a new probabilistic model for the evolution of double-stranded RNA molecules that considers substitutions of the base pairs rather than of each of the bases independently. The new model, called the double-stranded model, was incorporated into the neighbor-joining distance and maximum likelihood methods. Computer simulations show that maximum likelihood is very robust to the violation of the assumption of the independence of sites. In contrast, the neighbor-joining method is sensitive to such violations: the double-stranded model can provide a significant increase in the chance of obtaining the correct tree topologies with neighbor joining when distances are large and the tree is difficult to obtain. The new model also leads to lower but more realistic estimates for the statistical confidence in the branch lengths and tree topologies.

Introduction

Ribosomal RNA sequences have been extensively studied by molecular systematists as they are present in all forms of life and are for the most part conserved enough to allow the alignment and phylogenetic analysis of widely diverse organisms (see Hillis and Dixon [1991] and Olsen and Woese [1993] for review). The evolution of the rRNA genes is also thought to accurately reflect the evolution of the genome since they are unlikely to be subject to processes that can produce discrepancies between the phylogeny of the gene and that of the species that carries it (see Doyle [1992] for review). The large size of the large subunit (LSU) and small subunit (SSU) rRNA sequences present a large data set of characters (the bases in the sequence) that are considered to evolve independently. Such a large character set reduces the chance of phylogenetic analysis being misled by chance homoplasies and increases the reliability of the trees obtained.

Key words: compensatory substitutions, evolutionary model, RNA evolution, maximum likelihood, neighbor-joining, phylogenetic analysis, computer simulations, statistical tests.

Address for correspondence and reprints: Elisabeth R. M. Tillier, New York State Department of Health, Wadsworth Center for Laboratories and Research, David Axelrod Institute, P.O. Box 22002, Albany, New York 12201-2002.

Mol. Biol. Evol. 12(1):7–15. 1995.

0737-4038/95/1201-0002$02.00

The assumption of independence of the characters is more for convenience than a true reflection of reality. Most methods of phylogenetic analysis assume that characters (bases or sites) evolve independently. This is deemed necessary as the mathematical formulation of the methods would be greatly complicated if the probability of a substitution at one site were dependent on the identity of a base at another. The assumption of independence of characters is most clearly violated by RNA sequences, especially rRNA, and also tRNA and catalytic RNAs. These molecules form base pairs that give them their characteristic secondary and tertiary structures. The formation of these structures is necessary for the function of the molecule and is, for the most part, conserved throughout evolution (see, e.g., Gutell et al. 1992).

The results of phylogenetic sequence comparisons show that a large proportion of the sequence in the small subunit (SSU) rRNA molecule is base paired. An example of this was shown by Vawter and Brown (1993), who found that over 40% of the molecule consists of paired sequences. The proportion of paired sequences may be underestimated since loops, bulges, and other "single-stranded" regions may in fact be involved in tertiary interactions. Wheeler and Honeycutt (1988) have shown that the observed number of compensatory substitutions in 5S rRNA sequences was much higher than

would be expected if substitutions occurred independently. These authors suggested that the substitutions in base-paired regions should be weighted by a factor of one-half or be discarded altogether from the data. Since a large proportion of the molecule is paired, weighting would be preferable, but Dixon and Hillis (1993) recently argued that a weighting of one-half is too severe and showed how to obtain a more accurate weighting factor for use with maximum parsimony. Weighting may be appropriate when using the maximum parsimony method, but weighting is not appropriate when using model-based methods (maximum likelihood and some distance methods), since weighting does not correct for the inadequacies of the model. These methods will give statistically unsound and possibly erroneous results when the data do not meet the assumptions of the model used. All the models of substitutions in current use, such as the Kimura (1980) two-parameter model, have in common the fact that the probability of a substitution does not depend on the occurrence of substitutions at other sites in the molecule. Unfortunately, this is surely not a valid assumption for base-paired sequences, which means that, for these sequences, the current models of substitution are incorrect.

To determine whether the accuracy of the distance matrix and maximum likelihood methods is affected when the assumption of independence is not met by the sequences, we developed a model of substitution specific for double-stranded regions of RNA sequences that considers the substitution of base pairs rather than those of the individual bases. Distance equations as well as an implementation of the maximum likelihood procedure were obtained for this new model. This approach allows an examination of the accuracy of standard distance and maximum likelihood methods in estimating tree topology and branch lengths when the sequences have evolved following this "double-stranded" model. When sites do not evolve independently, the effect on any statistical test could be very serious since the number of independent data elements will be overestimated. Specifically, we have investigated the effect on a statistical test for the tree topology developed by Tajima (1992) that is used with the distance method. In addition, the potential effects of base pairing on other statistical tests used with maximum parsimony and maximum likelihood are also discussed.

Methods

The Model

To determine the effect of base pairing on the accuracy of the standard models that assume the bases in the sequence evolve independently (which we will call "single-stranded" models), it was necessary to develop a model for the evolution of base pairs in the RNA as a standard for comparison and for the evolution of these base-paired sequences in computer simulations. If we consider only base-paired regions of an RNA molecule, each pair can be considered as the evolutionary character. If, of all 16 possible base combinations, only Watson-Crick pairings are allowed (A-U and G-C), then the bases are 100% dependent on one another (all substitutions are double-base substitutions), and it would be necessary to consider only half of the bases in sequence. However, non-Watson-Crick interactions can also form in RNA sequences (Chastain and Tinoco 1991; Gutell et al. 1992), and of those, G-U base pairings are by far the most common (which make up 10%–25% of the base pairs) and also need to be considered.

A schematic representation of the model of substitution for paired sequences is shown in figure 1. We call this the double-stranded model. Its transition probability matrix was described in a previous paper (Tillier 1994). The model allows for only three possible base pairs: A-U, G-C, and G-U. All other base combinations are considered as single-stranded sites and not considered with this model. A more complete model that allows other base combinations will be reported later (E. R. M. Tillier and R. A. Collins, unpublished data). All transversion substitutions are double-base substitutions occurring at an instantaneous rate β. We assume that single transitions occur to and from the G-U base pair at an instantaneous rate α_s (only a single base is changed in this case). Although G-U base pairs are common in rRNA, some positions in the sequence never, or rarely, allow a G-U, which suggests a strong selection against G-U's at certain positions. Transition substitutions are, however, very common, even at these sites. Cons

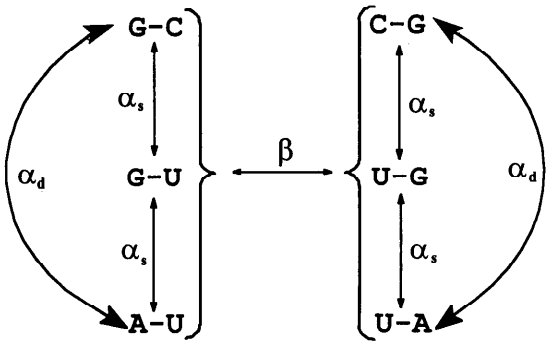

FIG. 1.—Model of substitution for RNA base pairs. This is a schematic representation of an instantaneous substitution model. There are six possible states for a site: A-U, G-U, G-C, U-A, U-G, and C-G. These can be considered in two groups, diagrammed here as the left and right columns. Within each group, 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. Transversion substitutions occur between these two groups at rate β. The α_s, α_d, and β are the rates of substitutions per unit time.

quently, a rate α_d of instantaneous double transitions (from A-U to G-C and vice versa) is allowed in the model. Note that β, α_s, and α_d are rates of *substitution* and are not meant to imply anything about mechanisms or rates of *mutation*. The base-pair frequencies are allowed any value but are assumed at equilibrium.

The equation for the modified distance measure between two sequences, separated by the evolutionary time $2t$ is obtained as in Cox and Miller (1977) from the transition probability matrix derived elsewhere (Tillier 1994*a*, 1994*b*) and is given as

$$K(2t) = -\frac{1}{2}[1-2\pi_{gu}(1-\pi_{gu})-2\pi_{au}\pi_{gc}]\ln(1-2Q)$$
$$-2\left[\pi_{gu}(1-\pi_{gu})-\frac{\pi_{au}\pi_{gc}\pi_{gu}}{1-\pi_{gu}}\right]\ln\left(1-Q-\frac{S}{2\pi_{gu}(1-\pi_{gu})}\right) \quad (1)$$
$$-2\frac{\pi_{au}\pi_{gu}}{1-\pi_{gu}}\ln\left(1-Q-\frac{S}{2(1-\pi_{gu})}-D\frac{1-\pi_{gu}}{2\pi_{au}\pi_{gc}}\right).$$

Q, S, and D represent the frequencies of transversions and single and double transitions observed between the two sequences, respectively, and π_{au}, π_{gu}, and π_{gc} are the equilibrium frequencies of the base pairs. The expression for the sampling variance of the distance was obtained according to Kimura (1980) from

$$\sigma_K^2 = E[(\sigma K)^2], \quad (2)$$

where δK is a small change in K given by

$$\delta K = \frac{\partial K}{\partial S}\delta S + \frac{\partial K}{\partial D}\delta D + \frac{\partial K}{\partial Q}\delta Q \quad (3)$$

and where $\partial K/\partial S$, $\partial K/\partial D$, and $\partial K/\partial Q$ are the derivatives of K with respect to S, D, and Q that can be obtained from equation (1). We also have

$$E[(\delta Q)^2] = \frac{Q(1-Q)}{N}, \quad E[(\delta S)^2] = \frac{S(1-S)}{N},$$
$$E[(\delta D)^2] = \frac{D(1-D)}{N}, \quad E[\delta Q\delta S] = -\frac{QS}{N}, \quad (4)$$
$$E[\delta Q\delta D] = -\frac{QD}{N}, \quad E[\delta D\delta S] = -\frac{DS}{N},$$

where N is the number of sites. The distance formula is used to compute the matrix of pairwise distances for input into the neighbor-joining algorithm (Saitou and Nei 1987; Studier and Keppler 1987) to obtain a phenogram. An implementation of the maximum likelihood procedure for this model has been also described previously (Tillier 1994*b*).

This double-stranded probability model is similar to previous single-stranded substitution models (such as the Kimura [1980] two-parameter model), but it considers the base pairs as the characters rather than the bases themselves. Otherwise, methods that use the double-stranded model will make the same assumptions as with single-stranded models (i.e., the characters evolve independently and at the same rate according to a stationary Markov process; Cox and Miller 1977).

Simulations

With a model in place, a possible way to determine the effect of pairing could be to consider RNA sequences from species with known phylogenies. Such sequences could be analyzed with different models and methods in order to determine whether the double-stranded model gives a better estimate of the known solution than single-stranded models. Such an analysis would require several independent phylogenies in order to determine whether an improvement is statistically significant. Unfortunately, there is no available data set of known phylogenies in which both topologies and branch lengths are known.

We have chosen an alternative approach using computer simulations to provide known, testable phylogenies. Computer simulations have been used extensively to determine the effectiveness of the different tree-building algorithms (Sourdis and Nei 1988; Saitou and Imanishi 1989; Jin and Nei 1990) and to determine the effect of violating the assumptions of models of substitution (Fukami-Kobayashi and Tateno 1991). This "in computo" approach has the advantage that evolution can be simulated under different conditions. Different values for the parameters in the model of substitution, different trees, different lengths of sequences, and other factors can be used to determine the robustness of the phylogenetic analysis methods to these varying conditions.

The evolution of sequences was simulated by making substitutions in a completely double-stranded random sequence of a given base-pair composition. Substitutions in the sequences occurred according to their probability, given the double-stranded model of substitution (fig. 1) and the known branch lengths of the tree in figure 2 rooted in the middle of the interior branch (as described in Tillier [1994*a*, 1994*b*]). Four sequences were thus obtained and then used to infer an unrooted phylogeny with both single-stranded and double-stranded models using neighbor joining and maximum likelihood. For neighbor joining, the Kimura (1980) distance measure was used for the single-

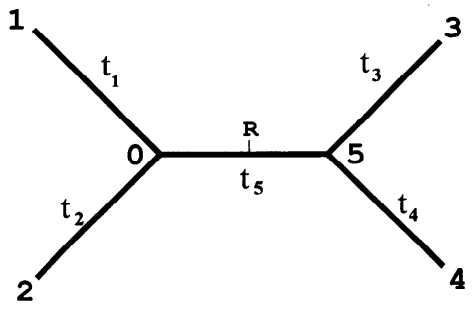

FIG. 2.—Generalized tree for four sequences. The nodes are numbered from 0 to 5. The node *R*, in the middle of the interior branch, denotes the placement of the root of the tree, the starting point for the simulations. The branch lengths t_1–t_5 correspond to the time (as measured in years or generations) between the two nodes connected by that branch.

stranded distance method, and the modified distance equation for the RNA model given by equation (1) was used for the double-stranded method. For maximum likelihood, the likelihoods of the three possible unrooted tree topologies were maximized using the EM algorithm, as derived previously (Tillier 1994*b*) for both a two-parameter, single-stranded model (Hasegawa et al. 1985) and the RNA double-stranded model described here. The single-stranded model allows different base frequencies and different rates of transitions to transversions (which need not be specified). When applied to the double-stranded RNA sequences, the single-stranded model is simplified since, for these sequences, the frequency of purines necessarily equals that of pyrimidines. The Hasegawa et al. (1985) single-stranded model was used in comparison with our double-stranded model because it most resembles our double-stranded model in that it makes only one additional assumption: that the bases evolve independently.

Besides the simple determination of the number of times the correct tree was obtained in the simulations, a measure of the discrepancy between the estimated and actual pairwise distances in the tree, S_r (Tateno et al. 1982), was also calculated and given by

$$S_r = \sqrt{\frac{2\sum_{j<i}^{n}(d_{ij}-d'_{ij})^2}{n(n-1)}} \qquad (5)$$

where n is the number of sequences (four, in this case), d_{ij} is the normalized estimated distance between sequences i and j in the calculated tree, and d'_{ij} is the corresponding normalized distance in the tree used for the simulation (the true tree). The distances were normalized by dividing them by the estimated distance between sequences 1 and 3, to allow comparisons between models. Values for the mean S_r and its variance were also obtained for all these simulations.

Statistical Test

Tajima (1992) recently developed a statistical test for the tree topology similar to that of Nei et al. (1985) except that it did not depend on the assumption of a molecular clock. The test considers the quantity t, the estimated branch length of the interior branch (t_5^{est} is the

Table 1
Performance of the Double- (ds) and Single-Stranded (ss) Models in Obtaining the Correct Tree

				Neighbor Joining		Maximum Likelihood	
e[a]	f[a]	K_d[b]	K_s[b]	ss	ds	ss	ds
Tree A:[c]							
1 . . .	1	0.10	0.09	172	171	169	171
		0.44	0.40	176	185	176	180
		0.78	0.69	167	173	174	173
	3	0.10	0.08	180	178	171	173
		0.48	0.40	181	179	182	178
		0.80	0.65	163	162	179	180
0 . . .	1	0.10	0.09	166	169	171	176
		0.42	0.37	180	180	188	183
		0.75	0.65	155	164	173	171
	3	0.10	0.08	174	166	174	168
		0.44	0.33	181	172	183	178
		0.76	0.55	166[d,e]	140	171	172
Tree B:[c]							
1 . . .	1	0.10	0.09	118	121	107	113
		0.44	0.40	121	126	147	149
		0.78	0.69	121	135	151	148
	3	0.10	0.08	126	117	107	106
		0.48	0.40	110	126	155	157
		0.80	0.65	107	135[d]	139	146
0 . . .	1	0.10	0.09	120	119	98	105
		0.42	0.37	122	134	156	157
		0.75	0.65	102	121	144	139
	3	0.10	0.08	110	109	104	107
		0.44	0.33	109	116	156	160
		0.76	0.55	95[e]	114	131	139

NOTE.—Results are the number of correct trees obtained from 200 simulations on double-stranded sequences of 300 base pairs. The equilibrium base-pair frequencies were set to $\pi_{au} = 0.25$, $\pi_{gu} = 0.15$, and $\pi_{gc} = 0.60$.

[a] The simulations were carried out for various ratios of the values for the instantaneous rates in the model of fig. 1: $e = \alpha_d/\alpha_s$ and $f = \alpha_s/\beta$.

[b] K_d and K_s correspond to the distances (number of substitutions divided by the length of the sequence) expected between sequence 1 and the root of the tree (R) as calculated with the double-stranded model and the single-stranded model, respectively.

[c] Tree A, $t_1/t_2 = 1$ and $t_1/t_5 = 5$. In Tree B, $t_1/t_2 = 10$ and $t_1/t_5 = 25$ (see fig. 2).

[d] Indicates a statistical improvement (increase) in that model over the other in obtaining the correct tree topology at the 95% level.

[e] Indicates a statistical improvement (reduction) in that model over the other in the mean S_r value at the 95% level (data not shown).

Table 2
Performance of the Double- (ds) and Single-Stranded (ss) Models in Obtaining the Correct Tree When Base-Pair Frequencies Are Equal

				Neighbor Joining		Maximum Likelihood	
e	f	K_d	K_s	ss	ds	ss	ds
1	1	0.10	0.08	121	123	100	108
		0.48	0.41	125	142	146	149
		0.71	0.61	123	142[a]	152	153
	3	0.10	0.08	116	120	109	105
		0.50	0.38	129	139	150	156
		0.75	0.56	123	127	139	141
0	1	0.10	0.08	122	119	102	112
		0.48	0.40	130	138	161	161
		0.72	0.60	105	131[a,b]	154	154
	3	0.10	0.07	126	125	103	110
		0.47	0.32	139	141	161	159
		0.78	0.53	110[b]	123	142	145

NOTE.—Simulations were obtained with Tree B as in table 1, but with equilibrium base-pair frequencies $\pi_{au} = \pi_{gu} = \pi_{gc} = 1/3$.

[a] Indicates a statistical improvement (increase) in that model over the other in obtaining the correct tree topology at the 95% level.

[b] Indicates a statistical improvement (reduction) in that model over the other in the mean S_r value at the 95% level (data not shown).

estimate of the branch length t_5 in fig. 2) of a four-species tree, divided by the standard deviation of the estimate:

$$t = \frac{t_5^{est}}{\sqrt{V(t_5^{est})}}. \quad (6)$$

The variance of t_5^{est} is a maximum approximation since an estimate of its actual value can only be derived for a very simple one-parameter model. This variance is approximated as

$$V_{max}(t_5^{est}) = \frac{1}{4}\sum_{i=1}^{4}\sum_{\substack{j=1 \\ j<i}}^{4} V(d_{ij}) - \frac{d_{ij}}{N}, \quad (7)$$

where $d_{ij}$is the estimated distance between the sequences i and j, and $V(d_{ij})$ is the estimated variance of that quantity derived from equation (2) for the double-stranded model and given by Kimura (1980) for the single-stranded model. The number of sites is indicated by N. Since t should be zero or negative for the wrong tree topology, the greater the value of t, the greater the confidence in the tree. To apply the test, the value for t is determined and declared significant if it is greater than a value c, which will depend on the desired level of confidence. An approximation to the distribution of c for the Kimura two-parameter model (Kimura 1980) was obtained by Tajima (1992) using simulations.

Results

Tree Topology and Branch Lengths

To determine the extent to which the standard methods of neighbor joining and maximum likelihood were affected in the accuracy of their estimates when their assumption that the bases in the sequence evolve independently was violated, we investigated the behavior of these methods on double-stranded sequences under various conditions. Several values for the relative branch lengths in the tree, the length of the sequences, the base-pair frequencies, and other parameters in the model such as the transition-to-transversion ratios were set in the simulations to generate sequences.

Two hundred simulations were carried out for each set of parameters and for several trees. The results of the simulations are given in tables 1–3, which show the number of correct tree topologies obtained with the single-stranded and double-stranded methods. The tables also show whether there was any significant difference between the calculated mean S_r values for the two methods. These quantities give a measure of the accuracy in the estimates of branch lengths. In table 1, the result is shown for Tree A and Tree B. These trees differ in their relative branch lengths t_1/t_5 and t_2/t_5 (see fig. 2). Also, K_d, the expected proportion of base-pair substitutions

Table 3
Performance of the Double- (ds) and Single-Stranded (ss) Models in Obtaining the Correct Tree with 1,000 Base-Paired Sites

				Neighbor Joining		Maximum Likelihood	
e	f	K_d	K_s	ss	ds	ss	ds
1	1	0.10	0.09	165	162	165	166
		0.44	0.40	161	174	185	183
		0.78	0.69	142	173[a]	190	190
	3	0.10	0.08	165	166	170	172
		0.48	0.40	146	165[a]	182	184
		0.80	0.65	128	163[a,b]	188	188
0	1	0.10	0.09	152	157	168	169
		0.42	0.37	154	163	186	186
		0.75	0.65	138	164[b]	185	179
	3	0.10	0.08	152	153	159	166
		0.44	0.33	159	168	184	192
		0.76	0.55	143[b]	157	178	187

NOTE.—Simulations were obtained with Tree B as in table 1, but with 1,000 paired sites.

[a] Indicates a statistical improvement (increase) in that model over the other in obtaining the correct tree topology at the 95% level.

[b] Indicates a statistical improvement (reduction) in that model over the other in the mean S_r value at the 95% level (data not shown).

(distance) between sequence 1 and the sequence at the root of the tree was varied. For comparison purposes, K_s, the corresponding distance in terms of expected single-base substitutions, is also given. The values for K_d and K_s were chosen so as to cover the range of variation in the sequences that is observed in SSU rRNA (E. R. M. Tillier and R. A. Collins, unpublished observation). The values for all the parameters were chosen in most cases to approximate what they appear to be in the current SSU rRNA database (as will be shown in a later paper).

Simulations were carried out for various ratios of the values for the instantaneous rates in the model of figure 1: $e = \alpha_d/\alpha_s$ and $f = \alpha_s/\beta$. The number of sites for these simulations was 300, and the base-pair frequencies were $\pi_{au} = 0.25$, $\pi_{gu} = 0.15$, $\pi_{gc} = 0.60$ in tables 1 and 3 or were all equal ($\pi_{au} = \pi_{gu} = \pi_{gc} = 1/3$) in table 2. In order to determine the effect of the base-pair composition, we performed simulations for one of the trees in which all base pairs were given equal frequencies (table 2). The base and base-pair compositions were estimated from the empirical frequencies when estimating the trees. To determine the effect of sequence length, simulations were also done with 1,000 base-pair sequences (table 3). A selection of these simulation results are presented graphically in figures 3 and 4. These figures show the

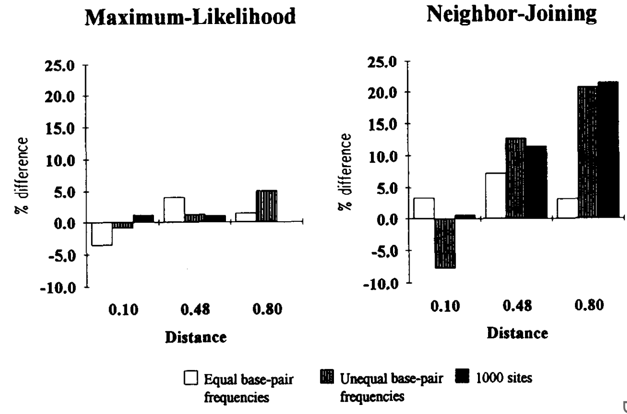

FIG. 4.—Difference in the ability of the double- and single-stranded methods of obtaining the correct tree from double-stranded sequences. The percentage difference between the number of correct trees obtained with the double- and single-stranded models relative to the double-stranded results for both the maximum likelihood and neighbor-joining methods is shown. Tree B was used for the simulations, and the difference is plotted against an increasing amount of expected divergence (distance) between the sequences. These are results selected from table 1 (unequal base-pair frequencies: $\pi_{au} = 0.25$, $\pi_{gu} = 0.15$, $\pi_{gc} = 0.60$; 300 paired sites), table 2 (equal base-pair frequencies: $\pi_{au} = \pi_{gu} = \pi_{gc} = 1/3$; 300 paired sites), and table 3 (unequal base frequencies: $\pi_{au} = 0.25$, $\pi_{gu} = 0.15$, $\pi_{gc} = 0.60$; 1,000 paired sites) with $e = 1$ and $f = 3$ in all cases.

reduction in the single-stranded models' ability to obtain the correct tree compared to that of the double-stranded model (with which the sequences were actually evolved and therefore expected to do better). A wider selection of trees and of parameter values have been explored (Tillier 1994*a*).

In most cases, there is no statistically significant difference in the observed mean S_r values between the single-stranded and double-stranded approaches with both the neighbor-joining and maximum likelihood methods. However, the results with the neighbor-joining method clearly show that there can be a significant reduction in the ability to obtain the correct tree with the single-stranded approach at higher sequence divergence (high K_d) and when the tree topology is "difficult" owing to a short interior branch length (small t_5) or some long exterior branches (high t_1). Tree A is not "difficult," and the single-stranded and double-stranded models perform just as well. However, in Tree B (table 1 and fig. 3), in which t_5 is 25 times shorter than t_1, there can be a reduction of up to 20% in the number of correct trees obtained with the single-stranded neighbor-joining method relative to the double-stranded neighbor-joining method.

The maximum likelihood method performs generally better than the neighbor-joining method, but, more important, it appears to be very robust toward the violation of the assumption that sites evolve indepen-

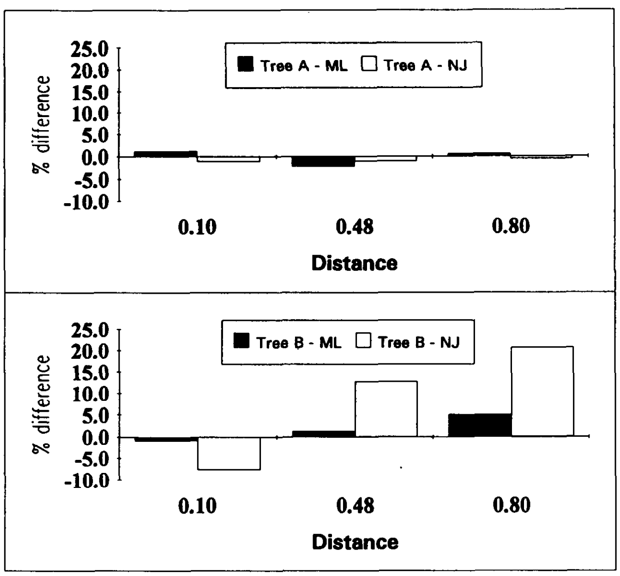

FIG. 3.—An example of the difference between the single- and double-stranded methods in their ability to obtain the correct tree. For two trees (A and B) and with both the maximum likelihood (ML) and neighbor-joining (NJ) tree-building algorithms, the differences between the number of correct trees found by the double- and single-stranded methods normalized as a percentage of the double-stranded results (% difference) is plotted against an increasing amount of expected divergence (distance) between the sequences. These results correspond to $e = 1$ and $f = 3$ in table 1.

dently. As is readily evident in the examples of figures 3 and 4, there is no significant reduction in the accuracy of the single-stranded maximum likelihood when used on double-stranded sequences compared to when the correct double-stranded model of substitution was used to estimate the tree. This finding was surprising since the likelihood of the tree is taken to be the product of the likelihoods at every site (Felsenstein 1981), which would, of course, only be true if the sites were independent.

Table 1 presents one case in which the double-stranded neighbor-joining method appeared to perform significantly worse than the single-stranded neighbor-joining method (Tree A, $e = 0, f = 3$) when the distance and the transition-to-transversion ratio were high. This was probably due to an overestimation of the rate of double transitions by the double-stranded model because a very high frequency of single transitions will lead to many apparent double transitions. In five of the 200 simulations, the tree was not even estimated as it was not possible to calculate at least one of the pairwise distances between the sequences, because a negative quantity was obtained in a term for which the logarithm is needed in the distance equation (1).

Tajima Test

To determine the difference in the behavior of the test statistic t in Tajima's test, 1,000 simulations were run with both the double-stranded and single-stranded neighbor-joining methods. For each tree estimated, it was determined whether it had the correct topology (i.e., the same as that used to obtain the sequences), and the value of the statistic t was calculated. Figure 5 shows the percentage of trees for which t was greater than a given value c. The top graph considers all the trees that were obtained with the incorrect topology, and the bottom graph shows this for those trees obtained with the correct tree topology. These graphs show that the distribution of c is wider for the double-stranded model than for the single-stranded model. For example, if $c = 1$, that is, if the tree topology is accepted as the correct one when $t > 1$, we will in fact accept the wrong tree a few times (<5%) with the double-stranded method, but more often (>10%) with the single-stranded method (see fig. 5*A*; $c = 1$). The single-stranded method thus has a greater probability of committing a Type I error (rejecting the correct tree topology and accepting an incorrect tree topology) than the double-stranded method. The right tree would also be accepted as correct more often with the single-stranded method (see fig. 5*B*; $c = 1$) so that the single-stranded method gives more confidence in the tree obtained, whether the topology was correct or incorrect.

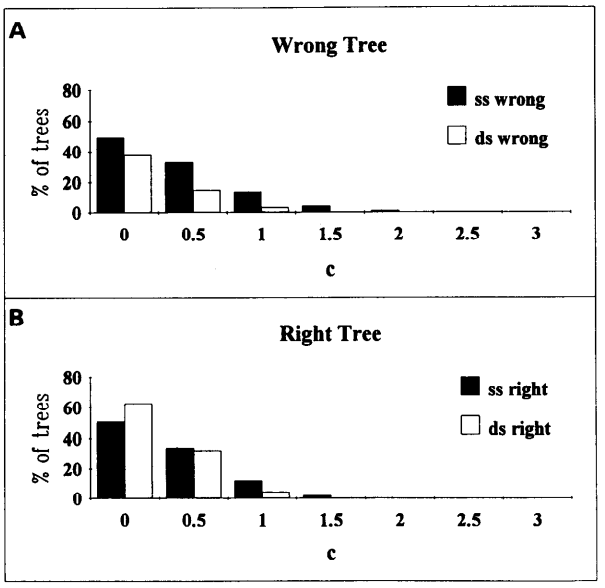

FIG. 5.—Distribution of Tajima's statistic with the single- and double-stranded models. The graphs show the percentage of correct and incorrect trees obtained with the double-stranded (ds) and single-stranded (ss) methods, out of 1,000 simulations, for which Tajima's statistic t (see text) was found to be greater than the value of c, plotted against increasing values of c. *A*, The percentage of those trees that had the incorrect tree topology is shown; *B*, the percentage of the trees obtained with the correct topology is shown. Tree B (see table 1) was used for these simulations, and the expected value for the distance in the interior branch t_5 was 0.02. Other values for the simulation were $\alpha_s/\beta = \alpha_d/\beta = 3$, $\pi_{au} = 0.25$, $\pi_{gu} = 0.15$, and $\pi_{gc} = 0.60$ over 300 sites.

Discussion

Previous model-based phylogenetic analyses of structured RNA sequences have assumed that all sites evolve independently of one another, an assumption clearly violated by conserved base pairings in these sequences. We have developed a model for the base-paired regions of RNA that takes into account the observed intersite dependence due to the maintenance of base pairing. The model allowed us to investigate by computer simulation the accuracy of the standard maximum likelihood and neighbor-joining methods in their ability to estimate the tree topology, branch lengths, and a measure of statistical confidence in the trees obtained when used with base-paired sequences. The accuracy of the standard neighbor-joining and maximum likelihood methods was compared to the accuracy of the corresponding methods that used the substitution model we derived for the evolution of the base pairs. This allowed us to determine the reduction in accuracy of the phylogenetic inference methods attributable to using the wrong model of substitution, as opposed to that expected from the greater sampling error that is inevitable because of the reduced number of independent characters.

Our simulations have shown that the maximum likelihood method is very robust toward the violation of the assumption that sites evolve independently made by the model of substitution. On the other hand, the simulations show that the neighbor-joining distance method can be sensitive to such violations, and its ability to obtain the correct tree topology can be significantly reduced because the sites are not evolving independently. This case results particularly when the tree is difficult and contains long branches. An investigation of the test statistic developed by Tajima (1992) reveals that the distance method will also overestimate the statistical confidence in the estimates of tree topology.

Actual RNA sequences will have a mixture of double- and single-stranded sites; therefore, the error expected from an inappropriate application of the single-stranded method will be reduced by the proportion of actually independent, single-stranded bases in the sequences. To combine both double- and single-stranded sites, it is necessary to know the structure in advance and for the two models to be applied separately to the appropriate sites. A combined likelihood function would simply be the product of the likelihoods at the two types of sites, while a combined distance measure would be the weighted average of the two distances. A difficulty would arise in such a combined analysis, since the double-stranded model of substitution proposed here allows only A-U, G-U, and G-C base pairs in the double-stranded regions. As shown by the SSU rRNA data set, many sites may sometimes have other base combinations even though there may be strong selection against unpaired bases in those regions. A more accurate and complete model that allows for other base combinations in double-stranded regions has been developed (E. R. M. Tillier and R. A. Collins, unpublished data).

The study of Tajima's (1992) test for tree topology led us to the observation that single-stranded methods give overestimates of the statistical confidence in the trees obtained from base-paired sequences. This result was expected since, in considering those double-stranded sites, the number of independent sites is reduced by one-half from the number of single-stranded sites, which thereby increases the variance of any estimate obtained from the sequences. This overestimation of the number of independent data will also affect statistical procedures that do not use a parametric formula for the variance of estimates, such as the bootstrap and the jackknife (see Felsenstein [1985] for review) or the one developed by Kishino and Hasegawa (1989) for use with maximum likelihood. A second, related reason for an unjustifiable increase in confidence is that, when a phylogenetic inference is made from an analysis of sequences containing both unpaired and paired sites, the patterns of substitution observed at the paired sites will be overrepresented in the data (and therefore in the bootstrap samples, for example). To apply statistical procedures to sequences containing both single- and double-stranded sites, it is therefore necessary to have an appropriate weighting factor for the double-stranded sites. A weighting scheme could be obtained by the method of Dixon and Hillis (1993). A weighting scheme based on probability models of substitution will be presented elsewhere (E. R. M. Tillier and R. A. Collins, unpublished data).

Our finding that the maximum likelihood method is robust to violation of independence of characters was a surprising and welcome result since it implies that, as long as single- and double-stranded sites are treated as separate categories, currently available implementations of maximum likelihood (DNAML and fastDNAML) will not be adversely affected (although confidence estimates will be overestimated). The neighbor-joining method is not robust, however, particularly when considering highly diverged sequences. The final outcome is that, when compared to a purely single-stranded analysis, taking base pairing into account can improve the chance of obtaining the correct tree but with a lower, more realistic confidence in the tree obtained.

Acknowledgments

We thank Tom Lew, Jeremy Carver, and the Department of Medical and Molecular Genetics for the use of their computers. We thank Pierre Potvin for his comments and suggestions on the manuscript. This work was supported by a research grant to R.C. from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Graduate Scholarships to E.T. from NSERC and the Ontario Graduate Scholarship program.

LITERATURE CITED

CHASTAIN, M., and I. TINOCO. 1991. Structural elements in RNA. Prog. Nucleic Acid Res. Mol. Biol. **42**:131–177.

COX, D. R., and H. D. MILLER. 1977. The theory of stochastic processes. Chapman & Hall, London.

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.

DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. Syst. Bot. **17**:144–163.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **30**:783–791.

FUKAMI-KOBAYASHI, K., and Y. TATENO. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitutions. J. Mol. Evol. **32**:79–91.

GUTELL, R. R., A. POWER, G. Z. HERZ, E. J. PUTZ, and G. D. STORMO. 1992. Identifying constraints on the higher-order structure of RNA: continued development and applications of comparative sequence analysis methods. Nucleic Acids Res. **20**:5785–5795.

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.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29**:170–179.

NEI, M., J. C. STEPHENS, and N. SAITOU. 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

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

SAITOU, N., and T. IMANISHI. 1989. Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum likelihood, minimum evolution and neighbor-joining models of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. **6**:514–525.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SOURDIS, J., and M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5**: 298–311.

STUDIER, J. A., and K. J. KEPPLER. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol. **5**:729–731.

TAJIMA, F. 1992. Statistical method for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. Mol. Biol. Evol. **9**:168–181.

TATENO, Y., M. NEI, and F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. **18**:387–404.

TILLIER, E. R. M. 1994*a*. Modelling the evolution of RNA secondary structure: implications for phylogenetic analysis. Ph.D. thesis, University of Toronto.

———. 1994*b*. Maximum likelihood with multi-parameter models of substitution. J. Mol. Evol. **39**:409–417.

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.

MITCHELL SOGIN, reviewing editor

Received May 12, 1994

Accepted August 24, 1994