Rates and Patterns of Base Change in the Small Subunit Ribosomal RNA Gene

Lisa Vawter[1] and Wesley M. Brown

Molecular Systematics Laboratory and Insect Division, Museum of Zoology, and Department of Biology, University of Michigan, Ann Arbor, Michigan 48109-1079

Manuscript received October 28, 1992
Accepted for publication February 4, 1993

ABSTRACT

The small subunit ribosomal RNA gene (srDNA) has been used extensively for phylogenetic analyses. One common assumption in these analyses is that substitution rates are biased toward transitions. We have developed a simple method for estimating relative rates of base change that does not assume rate constancy and takes into account base composition biases in different structures and taxa. We have applied this method to srDNA sequences from taxa with a noncontroversial phylogeny to measure relative rates of evolution in various structural regions of srRNA and relative rates of the different transitions and transversions. We find that: (1) the long single-stranded regions of the RNA molecule evolve slowest, (2) biases in base composition associated with structure and phylogenetic position exist, and (3) the srDNAs studied lack a consistent transition/transversion bias. We have made suggestions based on these findings for refinement of phylogenetic analyses using srDNA data.

THE accumulation of DNA sequence data from disparate taxa makes study of the nature of DNA evolution possible. Because certain classes of DNA characters tend to change in a related fashion, the nature of change within these classes can be explored statistically, so that we might learn about the evolution of molecules, as well as refine the assumptions for the use of molecular data in phylogenetic analysis.

A transition-transversion rate bias has been observed in primate mitochondrial DNA (BROWN *et al.* 1982). These authors noted that transitions (C ↔ T and A ↔ G changes) were more common than expected, given random change, and proposed that the mitochondrial DNA bias toward transitions is due to mutation bias. LI, WU and LUO (1985) found a transition bias in nuclear protein-coding genes and pseudo-genes, though the bias is not as pronounced here as in mitochondrial genes. This bias is often discussed theoretically in terms of the silent sites of protein-coding genes (*e.g.*, JUKES and BHUSHAN 1986; JUKES 1987). Because both BROWN *et al.* and LI *et al.* grouped all transitions and all transversions in their studies, they did not discuss specifically whether each transition, taken individually, was more common than each individual transversion. Though this does hold true of the BROWN *et al.* data set, whether it holds for the LI *et al.* data set is less clear (L. VAWTER, unpublished observations). However, except for those structural RNA genes in the mitochondrial genome (HIXSON and BROWN 1986), a transition/transversion bias has yet to be demonstrated in structural RNA genes.

Despite the lack of evidence bearing upon individual transition/transversion rates in structural RNA genes, phylogenetic analyses that require specific assumptions about them have been undertaken. Though MINDELL and HONEYCUTT (1990) did tally transitions and transversions for srDNAs, they did not calculate rates of change, and used taxa in their study that were too closely related to allow calculation of rates from their tallies because of sample sizes. A transition rate bias and equal transversion rates for different nucleotides have been used as assumptions for phylogenetic analysis of rDNAs using the method of invariants (LAKE 1988, 1989). The transition-transversion bias assumption has also been suggested for cladistic analyses of rDNAs (MISHLER *et al.* 1988; PATTERSON 1989; MICKEVICH and WELLER 1990). It is unclear how applicable a the transition-transversion rate bias assumption is for phylogenetic analysis using non-protein-coding genes, however, including srRNA (OLSEN and WOESE 1989; WOESE 1989).

We demonstrate a simple method to estimate relative rates of change among nucleotides for different structural classes within srDNA. This method has the advantage that it circumvents a constant molecular clock assumption. We then use these relative rates to evaluate the assumptions of a transition-transversion bias and equal transversion probabilities for srDNA.

METHODS AND RESULTS

The data set: Here, we present the set of srDNA sequence characters we used (nucleic acid positions),

[1] Current address: Museum of Comparative Zoology, Harvard University, Cambridge Massachusetts 02138.

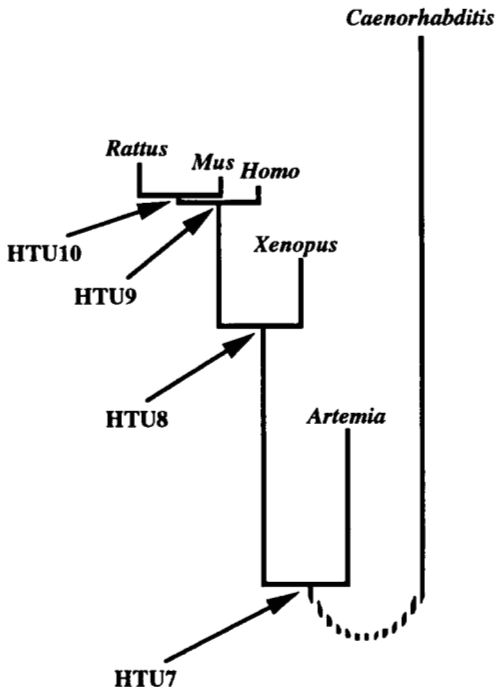

FIGURE 1.—Unrooted phylogenetic network of all OTUs and HTUs in the data set. The relationships among these taxa are noncontroversial (KEMP 1988; MILNER, 1988; NOVACEK, WYSS and MCKENNA 1988; WILLMER 1990). We used a parsimony method (HENNIG 1966), as implemented in the computer program PAUP, to assign base changes to branches. Branch lengths, excluding the dashed portion of the branch between Caenorhabditis and Artemia, are proportional to the number of changes they represent and are based on all characters, including cladistically noninformative characters. The consistency index (KLUGE and FARRIS 1969), based only on informative characters, is 0.94 and is discussed later in the text.

as well as the methods we used for inferring homology (in the evolutionary sense) among the characters. We used a parsimony method to infer hypothetical ancestral sequences for the evolutionary ancestors (HTUs) of the taxa that bear these srDNA characters. For this study, we used published srDNA sequences from a mouse, *Mus musculus* (RAYNAL, MICHOT and BACHELLERIE 1984), a rat, *Rattus norvegicus* (CHAN *et al.* 1984), a human, *Homo sapiens* (TORCZYNSKI, FUKE and BOLLON 1985), a frog, *Xenopus laevis* (SALIM and MADEN 1981), a brine shrimp, *Artemia salina* (NELLES *et al.* 1984) and a nematode, *Caenorhabditis elegans* (ELLIS, SULSTON and COULSON 1986). We chose these taxa because their evolutionary relationships are noncontroversial (Figure 1) (KEMP 1988; MILNER 1988; NOVACEK, WYSS and MCKENNA 1988; WILLMER 1990) and are derived from data sets other than srDNA. By doing this, we avoid the logical circularity that would result if we were to use srDNA sequence to derive an evolutionary network, and then derive conclusions about changes of srDNA sequence from the network derived from those same changes.

We inferred homology among the srDNA nucleic acid characters through alignment of both primary

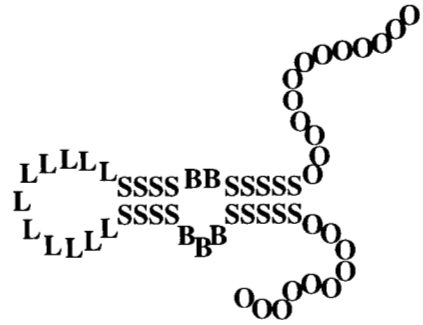

FIGURE 2.—Illustration of terminology for categorization of bases into bulges (*B*), loops (*L*), stems (*S*) and "other" (*O*). The "other" category comprises long single-stranded regions that are thought to interact with the ribosomal proteins (WOESE *et al.* 1983).

sequence and secondary structure. These alignments were performed as detailed in SOGIN and ELWOOD (1986), except that we discovered regions of sequence similarity by the method of LAWRENCE and GOLDMAN (1988), as implemented in EuGene 3.2 (Molecular Biology Information Resource, 1989). We confirmed the secondary structures from the literature using energetic (ZUKER 1989; JAEGER, TURNER and ZUKER 1990) and phylogenetic considerations (WOESE *et al.* 1983). Where the structure was ambiguous, we discarded the structural information. The aligned sequences will be provided electronically by the author (L.V.) upon request and are given in VAWTER (1991). The method of structural classification we used (stem, loop, bulge or "other") is shown in Figure 2. We emphasize here that we are including G-U pairs as stem structure, where they are not terminal to the stem. Unpaired regions within stems are classified as bulges, so that no unpaired positions are included as stem bases. The "other" class is not an arbitrary designation for positions with unknown structure; rather, it comprises long single-stranded regions that interact with ribosomal proteins (WOESE *et al.* 1983). All data, including those for which structure was ambiguous, were included in calculation of overall base compositions. Those positions where the alignment was unambiguous but the structure unknown (*e.g.*, positions 1226–1321 of the human sequence) were excluded from structural analysis. Sequence data that were not included in the remaining analyses because of ambiguity of alignment or structure are detailed in VAWTER (1991) and are available from the author (L.V.) upon request.

For the phylogeny in Figure 1, we calculated hypothetical ancestral sequences (HTUs) and predicted base changes for all varying positions using parsimony (PAUP; SWOFFORD 1989). As was pointed out by FITCH and MARKOWITZ (1970), it is necessary to superimpose sequence data on a phylogeny, rather than merely to tally differences between the taxa when taken pairwise, in order to utilize all changes required by the phylogeny. This approach to identifying base

	Rattus	*Mus*	*Homo*	*Xenopus*	*Artemia*	*Caeno-rhabditis*
Rattus	0.0	1.6	1.8	7.6	20.2	33.5
Mus		0.0	1.3	7.1	19.7	33.1
Homo			0.0	6.7	19.4	32.7
Xenopus				0.0	16.9	30.3
Artemia					0.0	24.6
Caenorhabditis						0.0

FIGURE 3.—Matrix of genetic distances between the taxa in the analysis. We calculated these as strict percent difference, with no corrections for multiple hits. As suggested by SOGIN and ELWOOD (1986), we omitted unique insertions from the calculations.

changes has the advantage over utilizing correction formulae to estimate numbers of changes between pairs of taxa, as it allows not only estimation of the number of changes that took place, but estimation of the specific changs that took place. Given a phylogenetic network and character states for the endpoints of that network, the algorithm in PAUP allocates character changes along that network so as to minimize its length. The prime assumption of this method is minimum evolution, which is reasonable in this case, because the small pairwise genetic distances (calculated as strict percent difference, not corrected for multiple hits) between these taxa (Figure 3) suggest that multiple changes at the same site are unlikely. The consistency index (KLUGE and FARRIS 1969) of this network (Figure 1, C.I. = 0.94, with cladistically noninformative sites excluded) suggests that multiple hits should not be a problem in this analysis or the analyses that follow. The consistency index is a measure of homoplasy that can vary between zero and one, with one indicating no homoplasy in the data set. Because they cannot be aligned, unique stretches of sequence [*e.g.*, positions 246–268 (human numbering system) in the mammals when compared to the rest of the data set] were excluded from the pairwise genetic distance calculations, as suggested by SOGIN and ELWOOD (1986).

We assigned base changes at nucleotide positions inferred to be homologous to appropriate branches of the phylogenetic network, and then calculated branch lengths as sums of changes assigned by the parsimony algorithm. In cases where the branch on which a change was inferred to have occurred was ambiguous, we divided the substitutional increment equally among the branches to which it could be assigned. We included all character changes in this analysis, including cladistically noninformative changes, because our aim was to study tendencies in character change statistically, rather than to derive a phylogeny.

All analyses were done twice, once with sequence and structural data from Caenorhabditis and the hypothetical ancestral taxa included, and once with them eliminated. Because of ambiguities in structure or alignment for the Caenorhabditis srDNA sequence, positions had to be excluded from it that were included in the other taxa. This alternate inclusion and elimination of Caenorhabditis from the analysis allowed comparison of results of the analyses performed on the same positions in all taxa with the results of the analysis performed with one of the data sets (Caenorhabditis) incomplete. Analyses were also done with the hypothetical ancestral taxa eliminated, because they are artificial constructs. Removing neither Caenorhabditis nor the hypothetical ancestral taxa from the analysis changed the results qualitatively, except for lowering the sample sizes. Therefore, the results of these additional manipulations will not be presented. The phylogeny shown in Figure 1 was imposed on the sequence data and is the foundation for all the analyses.

Structural composition: Analysis of structural composition allows comparison of relative rates of change in different structural classes. These relative rates can then be used to infer the importance of base sequence to the functions of the various structural classes. Structural composition of the data set is shown in APPENDIX A. Almost 43% of the total 11,340 nucleotide positions studied are involved in stem pairing. The "other" category (defined earlier) comprises about 20% of the positions. The bulge, loop and "unknown" categories were the smallest structural categories, comprising about 15%, 10% and 12% of the positions, respectively. The unknown category (positions for which structure could not be unambiguously assigned) comprised 9.8% and 13.7% of the data set in Artemia and Caenorhabditis, respectively. This reflects the finding that the *Escherichia coli* model of srRNA secondary structure (WOESE *et al.* 1983) may not fit arthropods and Caenorhabditis as well as it does other taxa (L. VAWTER and W. M. BROWN, unpublished results). The unknown category is largest in HTU 7 (almost

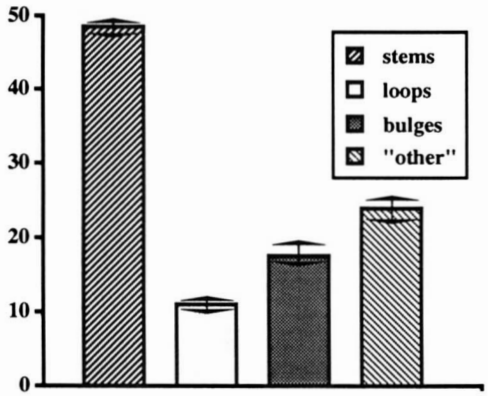

FIGURE 4.—The average percent of different structures, excluding unknown regions, for the extant taxa listed in APPENDIX A. The ranges of the structural compositions are indicated by the horizontal bars.

40% of the positions), because the genetic distance between Caenorhabditis and Artemia generates a large number of ambiguous character states (bases) in their hypothetical ancestor. These ambiguous character states are expected in any structural analysis that involves hypothetical ancestral taxa, because it cannot always be predicted which of two or more bases was more likely to be present at a position. These ambiguous character states frequently produce ambiguities in structure. For example, though it might be possible to predict that an ambiguous C or U would pair with G in a stem structure, it would be impossible to make a structural prediction for that same ambiguous C or U if it were opposite an A.

Figure 4 summarizes the means and ranges of structural compositions for the extant taxa studied, excluding positions in the unknown category. Excluding the unknown category gives a more accurate representation of the actual distribution of the structural classes of srRNA. The stem and loop classes are the most constant classes, in terms of proportion, with a range of less than 2%. The proportions allotted to the bulge and "other" categories vary a bit more, having ranges of 4.1% and 4.4%, respectively.

Base composition: Phylogenetic biases in base composition of genomic DNA (*e.g.*, BERNARDI and BERNARDI 1985; WADA, SUYAMA and HANAI 1991) and mitochondrial DNA (CLARY and WOLSTENHOLME 1985; CROZIER and CROZIER 1993) have been noted. Biases in base composition among srDNA structural classes might be predicted on the basis of energy considerations. For example, because the G-C pair has a lower free energy value than do A-U or G-U pairs (*e.g.*, FREIER *et al.* 1986; TURNER, SUGIMOTO and FREIER 1988), structural regions requiring base pairing might be expected to have a G/C base composition bias. Base compositions for the srDNAs studied, as well as of various structural components taken individually, are shown in Figure 5. Compositions are shown for each taxon separately, and for different structural regions of srRNA. As expected, stems are more G/C rich than are any of the other structural categories. Loop, bulge and "other" regions are much more A-rich than are stem regions. GUTELL *et al.* (1985) also noted that single stranded regions were A-rich and suggested that this might be because adenine is the least polar of the bases and thus might facilitate hydrophobic interactions with proteins. Regions of unknown structure do not differ in base composition appreciably from all structural classes combined (Figure 5). This is consistent with a lack of bias as to which structural components comprise the unknown class. Thus, it is unlikely that the inability to categorize unambiguously members of the unknown class biased the results of this analysis of base composition.

Phylogenetic biases in base composition are notable. Vertebrate srDNAs are more G/C rich overall [(G + C)/(A + T) = 1.24] than are those of non-vertebrate multicellular animals in this data set [(G + C)/(A + T) = 0.96]. In addition to the data presented in Figure 5, partial srRNA and srDNA sequences for other multicellular animals listed in GenBank were analyzed for base composition. In the data set presented in Table 1, it is notable that a cephalochordate, *Branchiostoma californiense* [(G + C)/(A + T) = 1.15], and an echinoderm, *Anisodoris nobilis* [(G + C)/(A + T) = 1.12], which are more closely related to vertebrates than the other non-vertebrates listed, also have a higher (G + C)/(A + T) ratio than the other non-vertebrate animals. Thus, using the other non-vertebrate taxa as outgroups, it might be inferred that this G/C-richness evolved prior to the evolution of vertebrates from the ancestral state of (G + C)/(A + T) ratio of approximately 1. Base composition in each of the structural classes is generally reflective of this phylogenetic base composition difference, with the only exceptional occurrence being the extraordinary T-richness and C-dearth for loop regions in Caenorhabditis srDNA.

Relative rates of change among structural classes: To compare relative rates of change among srDNA structural classes, we tabulated frequencies of change among the structural classes and then corrected those for the relative frequencies of the structural classes. Figure 6a shows a histogram of frequencies of change in the structural classes that comprise srDNA. The most frequent is clearly change in stem regions. However, when frequency of change is corrected for relative sizes of the structural classes (APPENDIX A), one can see that stem regions change no faster than other regions of the srDNA molecule on a per nucleotide basis (Figure 6b). Stem, loop and bulge regions change at about the same rate, with the "other" category (long single-stranded regions which presumably interact pri-

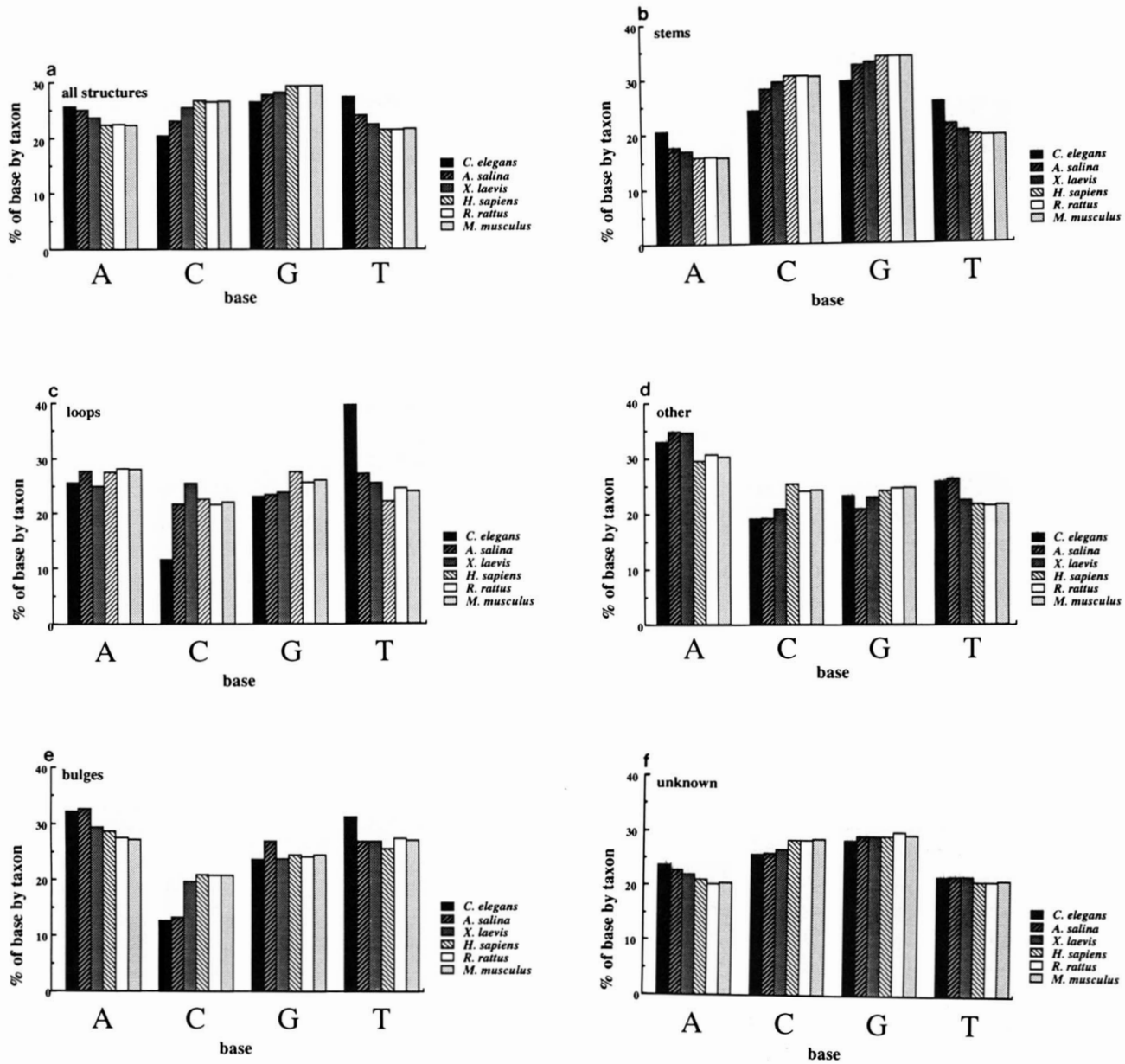

FIGURE 5.—(a–f) Histograms of base compositions of various structures of different taxa.

marily with the ribosomal proteins) evolving slowest. GUTELL *et al.* (1985) partitioned universally conserved nucleotides in srDNAs into those that occur in single-stranded (loop, bulge or other) as opposed to double-stranded (stem) regions. They noted that while only 39% of all nucleotides occur in single-stranded regions, 59% of universally conserved nucleotides occur in them. The finding of this study is consistent with the results of GUTELL *et al.* (1985) and suggests an interesting possibility. Because GUTELL *et al.* (1985) categorized nucleotides as single or double-stranded and did not consider the loop, bulge and "other" categories as distinct entities, the present study suggests that the distribution of universally conserved positions should be reexamined. Because stem and "other" categories comprise about 49% and 24% of the data set, respectively, it is possible that GUTELL *et al.*'s (1985) classification of srRNA positions as single- or double-stranded regions is an oversimplification. The "other" class might have numerically dominated the loop and bulge classes, incorrectly implying that single-stranded regions have a higher proportion of universally conserved nucleotides than double-stranded regions. That "other" regions evolve more slowly than the rest of the structural classes suggests that nucleotide sequence is critical in the function of these regions, perhaps in their interactions with ribosomal proteins or with the 5S or 28S rRNAs, in the tertiary structure of the 18S rRNA, or even in translation. A reexamination of the distribution of universally conserved positions might reinforce the idea of the importance of base sequence in "other" regions,

TABLE 1

(G+C)/(A+T) ratios for additional srDNA sequences of multicellular animals

Taxa	%A	%C	%G	%T	(G+C)/(A+T)
Tenebrio molitor insect (X07801)	24.6	23.3	27.4	24.7	1.03
Anisodoris nobilis echinoderm (M20097, M20098, M20099)	25.1	24.9	28.0	22.0	1.12
Bombyx mori insect (X01339)	25.2	23.1	28.2	23.5	1.05
Branchiostoma californiense cephalochordate (M20044, M20045, M20046)	24.9	24.5	29.1	21.6	1.15
Chaetoperus sp. polychaete (M20103, M20104, M20105)	26.8	21.2	27.2	24.9	0.94
Spisula solidissima bivalve (M20122, M20127, M20113)	25.4	23.7	28.3	22.6	1.08
Limulus polyphemus chelicerate (M20083, M20084, M20085)	27.2	23.2	27.3	22.3	1.02
Dugesia tigrina planarian (M20068, M20069, M20070)	30.7	18.5	23.6	27.2	0.73
Hydra sp. cnidarian (M20077, M20078, M20079	28.4	18.7	25.8	27.1	0.80
Lingula reevi brachiopod (M20086, M20087, M20088)	26.3	20.7	27.7	25.3	0.94

The numbers in parentheses following the taxon names are Genebank accession numbers for the sequences analyzed. Of these sequences, only the *T. molitor* sequence is a complete sequence.

and it might suggest that certain subsets of the loop and bulge regions are important in these interactions as well.

Base change matrices: To examine whether the finding of a transition-transversion bias in protein-coding genes (BROWN *et al.* 1982; JUKES and BHUSHAN 1986; JUKES 1987) can be extended to structural RNA genes, we calculated relative rates of base change for the srDNA data set overall, as well as for the different structural classes. We chose to calculate relative, rather than absolute, rates of change because a calculation of absolute rates requires estimates of time since divergence of taxa, a quantity that is rarely known accurately. Here, we present a method for estimating relative rates of base change that does not require constant rate assumption. We calculated matrices of base change in the following manner.

Step 1: Letting X and Y represent any two of the four nucleotides (A, C, G and U), we combined $X \rightarrow Y$ and $Y \rightarrow X$ state changes into a single category of $X \leftrightarrow Y$ state changes. This was necessary because changes are inferred from an unrooted network. Because of this, the polarity of base change cannot be inferred

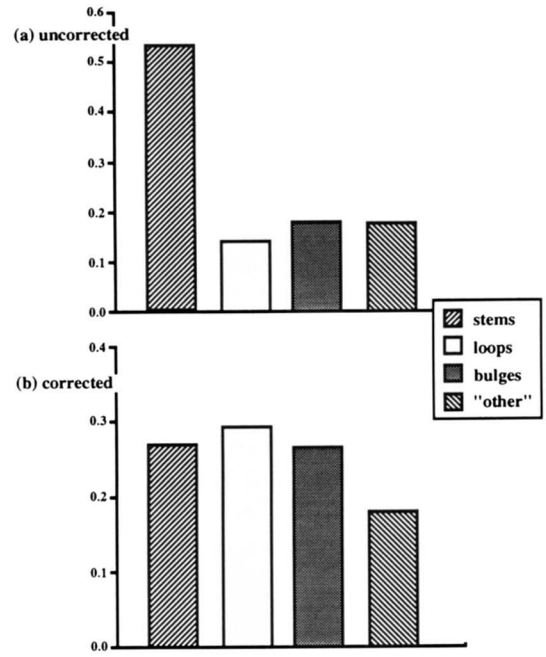

FIGURE 6.—(a) Histogram of frequencies of change among structural classes in srDNA. (b) Histogram of frequencies of change among structural classes in srDNA, corrected for relative frequencies of structural classes.

from the network, except in the case of changes on terminal branches.

For each of the pairs of taxa, these counts of base change: $\mathbf{N}_{\mathbf{X}\leftrightarrow\mathbf{Y}}$, where ($\mathbf{N}$ = number of changes between bases $\mathbf{X}$ and $\mathbf{Y}$; $\mathbf{X}$ = A, C, G, T; $\mathbf{Y}$ = A, C, G, T) were organized as 4 × 4 matrices. The raw base change counts for each of the pairs of taxa and for each structural class are presented in VAWTER (1991) and can be obtained from the author (L.V.). Base change counts in the structural categories are not always whole numbers. This is because a base change may cause the structural categories to be different in each of the taxa. For example, a G → C change may cause a U-G stem pair to become a U C bulge. In such cases, a half-count was added to each of the appropriate structural categories for that particular base change.

Step 2: Various structural classes and taxa had different base composition biases. Because stems, for example, are G/C rich, G and C are inherently more likely to be involved in stem base changes. Likewise, because Caenorhabditis and Artemia are more A/T rich than are the vertebrates in this data set (Figure 5), more changes between A and T are expected in Caenorhabditis and Artemia than in the vertebrates. We corrected counts of character state changes for the probability of finding each base in a particular structural category and pair of taxa, so that base compositions of the different taxa did not affect estimates of relative rates of base change (see Figure 7 for example). The correction factors were scaled to

Artemia x Caenorhabditis (stems)
X->Y and Y->X changes combined

uncorrected for base compositions ⟶ corrected for base compositions

	A	C	G	T
A	--	21	37.5	27.5
C	--	--	33.5	46
G	--	--	--	34.5
T	--	--	--	--

	A	C	G	T
A	--	26.2	39.6	38.0
C	--	--	25.7	46.2
G	--	--	--	29.4
T	--	--	--	--

FIGURE 7.—Examples of matrices showing tallies of raw base changes and base change tallies corrected for base composition as described in the text.

1.000. A correction factor of 1.000 would indicate that the bases involved in the base change under examination (*e.g.*, A and G in an A ↔ G change) each comprised 1/4 of all bases. For example, we observed 33.5 C ↔ G stem base changes between Artemia and Caenorhabditis. The correction factor by which the tally of changes was divided, 1.511, is reflective of the overrepresentation of both C and G in stem regions. The corrected figure, 25.7 C ↔ G changes, was then used in further calculations. The counts of character state changes and table of correction factors (VAWTER 1991) are available in electronic form from the author (L.V.).

Step 3: For all pairs of extant taxa, we graphed the numbers of changes that occurred among members of a structural class against numbers of changes between C and U, the most common type of change in each of the structural categories. In these and all further manipulations, numbers of changes used are those which have been corrected for structural and phylogenetic base composition biases. For structures where various taxa had different base composition biases, it was essential that the numbers of changes represented by each point be corrected, as described in Step 2, above, for the graphs to be linear. An example of graphed points, both with and without this correction, is presented in Figure 8. In the example, the correlation coefficient of the uncorrected points is 0.50; the correlation coefficient of the corrected points is 0.82. When base composition bias exists, as in the case of loop structures, correction for base composition biases improves estimates of relative rates. Each graphed point of Figure 8 represents the number of C ↔ G changes and the number of C ↔ T changes in loop regions for a particular pair of taxa. The slope of the graphed line is therefore the probability of a C ↔ G character state change, relative to the probability of a C ↔ T change. Graphs and equations for the complete data set are presented in Figure 9, a–e. In each case, the slope of the graph is the rate of a particular base change, relative to the rate of C ↔ T change.

This method of calculating relative rates has the advantage that a statistic (MANTEL 1967; SMOUSE, LONG and SOKAL 1986) can be employed to test the statistical significance of the relationships among the

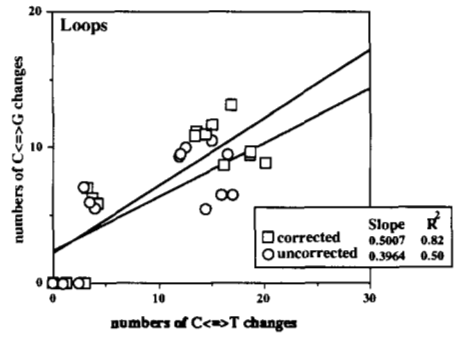

FIGURE 8.—Example of graphed points, both with and without the correction for base composition. As shown by the *R* values, the linearity of the relationship between the rate of C ↔ G and C ↔ T changes is improved by the correction.

base change pairs. The Mantel test is used to assess the significance of the association between two matrices, where the elements of each matrix may not be independent of the the other elements of the matrices. For the Mantel test, a Monte Carlo null distribution for the association of the two matrices is estimated as follows. One matrix is held constant, while the rows and columns of the other are permuted many times (n = 10,000, in this case). The distribution of the level of correspondence between the constant matrix and each of the 10,000 permutations (partial correlation coefficient) is plotted. To assess the significance of the association between the original two matrices, the level of correspondence (correlation coefficient) between them is assessed against this distribution. In the case of the srDNA data, the [C ↔ T] matrix was held constant and the relationship between it and each of the {[A ↔ C], [A ↔ G], [A ↔ T], [C ↔ G], and [G ↔ T]} matrices was assessed by comparison of the distribution of the correlations between it and the permuted versions of these matrices. P values for the comparisons of the [C ↔ T] matrix and each of the rest of the matrices were calculated from the Monte Carlo null distributions, and are given in Figure 9, a–e. With one exception, $P < 0.01$ for the comparisons. In the case of the comparison of the [A ↔ G] matrix with the [C ↔ T] matrix for loops, $P < 0.05$. This P value is perhaps influenced by the small sample size (n = 61.5) of A ↔ G changes in loops. Statistical assessment is important because in any comparative study, the hierarchical nature of a phylogeny prevents the character states that have evolved along the various phylogenetic paths from being independent of each other (FELSENSTEIN 1985). In other words, because changes in lineages accrue with time, those lineages that have been separated for long periods of time will tend to be less similar than those that have been separated for short periods of time. For example, Rattus and Homo are more likely to share any character state with each other than either is with Xenopus, simply because Rattus and Homo share a longer

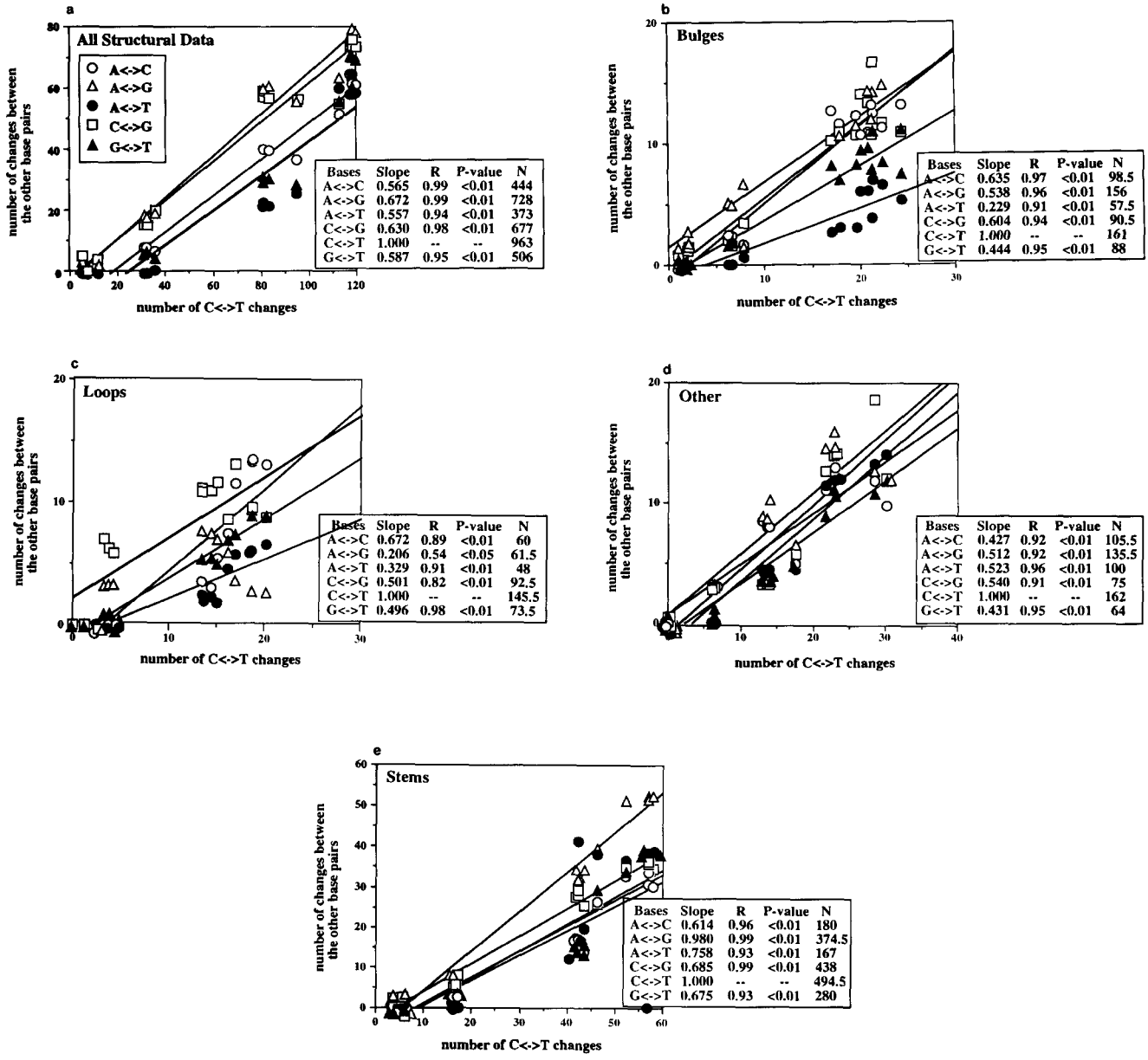

FIGURE 9.—(a–e) In these graphs, the points on a line represent the numbers of corrected base changes for each of the pairs of extant taxa for one of the $X \leftrightarrow Y$ base change pairs (where $X \leftrightarrow Y$ = A ↔ C, A ↔ G, A ↔ T, C ↔ G, or G ↔ T), graphed against corrected numbers of C ↔ T changes, for those pairs of taxa. In these graphs, when several points lay directly on top of one another, each was moved very slightly such that all might be visible. Slopes, R values, and statistical significances of the lines, using the Mantel test (MANTEL 1967; SMOUSE, LONG and SOKAL 1986) are given.

common evolutionary history with each other than either does with Xenopus (Figure 1). Thus it is helpful to be able to test the statistical significance of relationships between variables in a comparative study, because of the non-independence of points caused by some organisms in the study being more closely related to each other than to others.

In the manner presented above, one can estimate relative probabilities of change between all pairs of taxa for a particular class of characters without knowing times since divergence and without making a constant rate assumption. The use of a time variable is often impractical because time since divergence is generally not known accurately. More importantly, graphing change against time since divergence assumes a constant rate and, as is clear from Figure 1, this assumption is inappropriate in the case of these srDNAs.

Relative rates of base change: To examine the validity of a transition-transversion bias assumption for srDNA, we estimated relative rates of base change in the different structural categories and overall. We estimated relative rates as described above, in Step 3. The rate of C ↔ T changes was arbitrarily designated as 1.000. These relative rates are shown in Table 2 and are summarized in Figure 10. Across all structural

TABLE 2

Relative rates of change by structural category

Structure	A ↔ C	A ↔ G	A ↔ T	C ↔ G	C ↔ T	G ↔ T
All	0.565	0.672	0.557	0.630	1.000	0.587
Bulges	0.635	0.538	0.289	0.604	1.000	0.444
Loops	0.672	0.206	0.329	0.501	1.000	0.496
Other	0.427	0.512	0.525	0.540	1.000	0.431
Stems	0.614	0.980	0.758	0.685	1.000	0.675

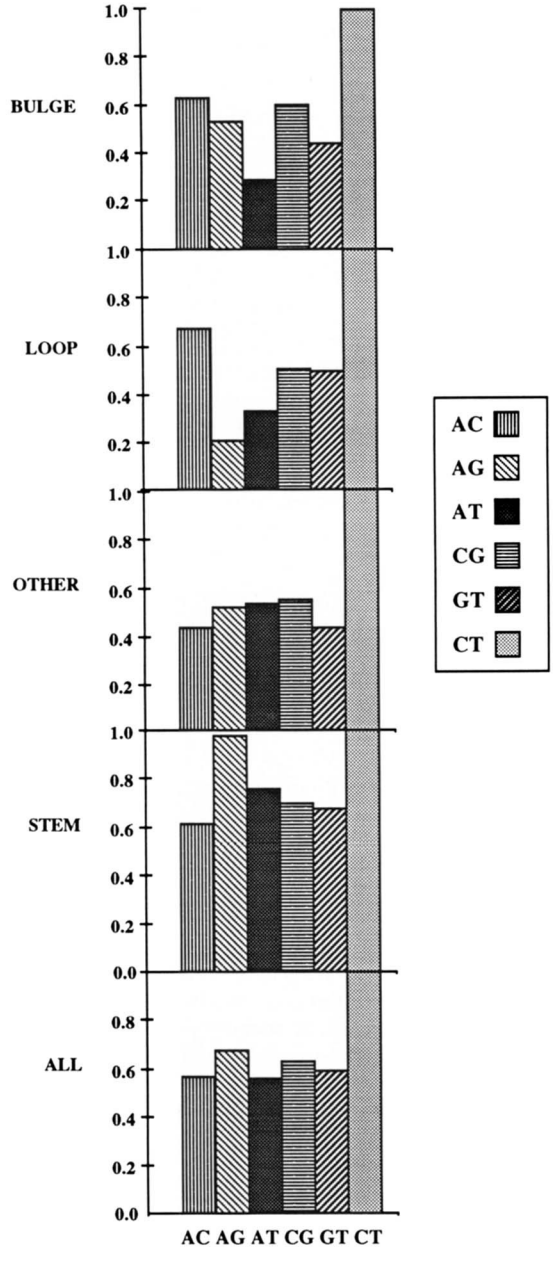

FIGURE 10.—This histogram illustrates relative rates of base change in the different structural categories as well as overall.

categories, C ↔ T changes, which are transitions, occur at the greatest rate. The other transition, G ↔ A, occurs at about the same rate as the remaining base changes, when all structural categories are considered together. However, amount of change between G and

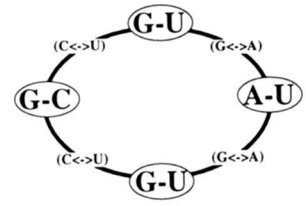

FIGURE 11.—Illustration of the single base changes which allow maintenance of base pairing in structural RNAs, though not DNAs. In structural RNAs, stems can be preserved with certain single base changes, because of G-U pairing.

A varies widely, from 20.8% of all changes in stems to only 6.4% in loops. The C ↔ T transition occurs at the highest rate of any change in all of the structural classes, as well as in all structural classes taken together. Though the A ↔ G transition occurs at almost the same rate as the C ↔ T transition in stems, it is actually the least common change in loops, and there are transversions that occur at a higher rate than the A ↔ G transition in all of the rest of the structural classes. Evidence also exists for a C ↔ T bias in amniotes (MARSHALL 1992), but no formal analysis was undertaken. The increased rate of A ↔ G transitions in stems is consistent with selection for maintenance of base pairing in stem structures, as illustrated in Figure 11. In DNA, where the complementary base pairs are C-G and A-T, it is not possible for base pairing to be maintained with only a single base change. For example, a C-G pair requires two base changes to become a G-C pair, an A-T pair, or a T-A pair. However, in RNA, where uracil (U) is substituted for thymine (T) and the G-U pair is stable, certain single changes in srRNA stems allow stem structure to be maintained (Figure 11). The effects of selection on base change in srRNA stem structure have been explored in detail (L. VAWTER and W. M. BROWN, unpublished data). It is clear that a transition-transversion bias does not hold for srDNA. Thus, a transition bias assumption, *per se*, should not be made for phylogenetic analysis of srDNA data. However, though it would require much structural analysis and a large investment of computer time, the relative rates of change in the various structural classes could be employed in phylogenetic analysis using either the method of invariants (LAKE 1988) or a maximum likelihood analysis. Also, this relative rate information could be used as ranking criteria for statistical comparisons of alternate tree topologies using the method of TEMPLETON (1983).

To examine the validity of assumption of equal transversion probabilities used for phylogenetic analysis of rDNAs with the method of invariants (LAKE 1988, 1989), we compared rates of transversions in the different structural regions as well as overall. Across structural categories, the rate of C ↔ G changes is most constant, ranging across a factor of

only 1.36. The rate of A ↔ T changes is least constant, varying by a factor of 2.62-fold. Though this may be interesting, it does not, however, address the validity of a constant transversion rate assumption. Within a structural category, the rate of transversions was least variable in the stem and "other" categories, with the highest rate of change being approximately 1.2 times more common than the lowest rate of change. Bulges and loops had, respectively, a highest rate of change 2.2 and 2.6 times that of their lowest rates of change. However, only a small portion of the bases in srDNA take part in loop and bulge structures (APPENDIX A, Figure 4). Overall, with the bases classed as structurally unknown included, transversions vary only 1.1-fold from the highest to the lowest rate. Thus the assumption of equal rates of transversions made in applying the method of invariants to srDNA is most likely valid.

Except for stem regions, we can find no particularly convincing rationalization for the patterns of base change we observed. We note that the most common base changes in stem regions, C ↔ T and G ↔ A, are those base changes which are expected to be favored under a regime of selection for maintenance of srRNA base pairing structure (L. VAWTER and W. BROWN, unpublished results). It is those specific changes which allow G-U pairs to be formed and base-paired structures to be maintained when particular members of C-G and A-U pairs change.

Different rates of base change are also apparent among the different taxa. The result of assigning nucleotide changes to branches of the phylogenetic network is shown in Figure 1. Because the allocation of changes was done using cladistic methodology and including all changes, as opposed to only cladistically informative changes, branch lengths for sister taxa in the rooted part of the network would be approximately the same if a constant rate of substitution held for this data set. Instead, it is apparent that a constant rate assumption is inappropriate for this set of srDNAs. High rate variation is expected in rDNAs when gene family size varies among taxa (OHTA 1983), as it does in this group of taxa (LEWIN 1987). This illustration of rate variation emphasizes the importance of verifying an assumption of rate constancy if it is to be used in phylogenetic analysis.

CONCLUSIONS

We have developed and applied a method of analysis for relative rates that is simple, standardizes for taxonomic and structural biases in base composition, and that avoids a constant rate assumption. The examination of relative rates of change of different categories of nucleotide in the srRNA gene permits the following conclusions.

First, stem, loop and bulge regions evolve at about the same rate. The long, single-stranded regions that presumably interact with proteins evolve slowest.

Second, there are structure-associated biases in base composition. Stems are more G/C rich than are any of the other structural categories, presumably because the G-C pairing confers maximum thermodynamic stability. Loop and "other" regions, which are suspected of interacting with proteins, are much more A-rich than are stem and bulge regions. This is in accord with the suggestion of GUTELL *et al.* (1985) that A-richness of these structures might facilitate interactions with proteins.

Third, there is also phylogenetic bias in base composition. Vertebrate srDNAs were much more G/C rich than were those of invertebrates (G + C/A + T of 1.24 for vertebrates, 0.96 for invertebrates), with base compositions of the various structural categories being generally reflective of this overall bias.

Fourth, a consistent transition-transversion bias does not exist in these srDNAs. Indeed, the relative rates of the various transitions and transversions vary more than fourfold among all structural categories. Though the C ↔ T transition occurs at a higher rate than do the other base changes, the other transition, A ↔ G, varies from being quite a bit more common than transversions in stems to being the least common change in loops. The relative rate scheme suggested in this manuscript is a more appropriate set of assumptions than the transition-transversion bias for phylogenetic analysis using the method of invariants, maximum likelihood, or the method of statistical inference suggested by TEMPLETON (1983). The results presented here call into question the validity of conclusions of phylogenetic analyses using the assumption that srDNA shows a transition-transversion bias.

Last, transversions vary only 1.1-fold from the highest to the lowest rate. Thus the assumption of equal rates of transversions made in applying the method of invariants to srDNA is probably a valid assumption.

The authors are grateful to C. W. BIRKY, B. CRESPI, D. FISHER, W.-H. LI, B. O'CONNOR, M. SLATKIN, P. SMOUSE and two anonymous reviewers for helpful suggestions. This research was supported by grants from the University of Michigan, the National Science Foundation and the U.S. Public Health Service, and was done was in partial fulfillment of the requirements of the Ph.D. at the University of Michigan by L.V.

LITERATURE CITED

BERNARDI, G., and G. BERNARDI, 1985 Codon usage and genome composition. J. Mol. Evol. **22:** 363–365.

BROWN, W. M., E. M. PRAGER, A. WANG and A. C. WILSON, 1982 Mitochondrial DNA sequences in primates: tempo and mode of evolution. J. Mol. Evol. **18:** 225–239.

CHAN, Y.-L., R. GUTELL, H. F. NOLLER and I. G. WOOL, 1984 The

nucleotide sequence of a rat 18S ribosomal ribonucleic acid gene and a proposal for the secondary structure of 18S ribosomal ribonucleic acid. J. Biol. Chem. **259:** 224–230.

CLARY, D. O., and D. R. WOLSTENHOLME, 1985 The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. J. Mol. Evol. **22:** 252–271.

CROZIER, R. H., and Y.-C. CROZIER, 1993 The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. Genetics **133:** 97–117.

ELLIS, R. E., J. E. SULSTON and A. R. COULSON, 1986 The rDNA of *C. elegans*: sequence and structure. Nucleic Acids Res. **14:** 2345–2364.

FELSENSTEIN, J., 1985 Phylogenies and the comparative method. Am. Nat. **125:** 1–15.

FITCH, W. M., and E. MARKOWITZ, 1970 An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4:** 579–593.

FREIER, S. M., R. KIERZEK, J. A. JAEGER, N. SUGIMOTO, M. H. CARUTHERS, T. NEILSON and D. H. TURNER, 1986 Improved free-energy parameters for predictions of RNA duplex stability. Proc. Natl. Acad. Sci. USA **83:** 9373–9377.

GUTELL, R. R., B. WEISER, C. R. WOESE and H. F. NOLLER, 1985 Comparative anatomy of 16s-like ribosomal RNA. Prog. Nucleic Acid Res. Mol. Biol. **32:** 155–216.

HENNIG, W., 1966 *Phylogenetic Systematics*. University of Illinois Press, Urbana.

HIXSON, J. E., and W. M. BROWN, 1986 A comparison of small ribosomal RNA genes from the mitochondrial DNA of great apes and humans: sequence, structure, evolution and phylogenetic implications. Mol. Biol. Evol. **3:** 1–18.

JAEGER, J. A., D. H. TURNER and M. ZUKER, 1990 Predicting optimal and suboptimal secondary structure for RNA. Methods in Enzymol. **183:** 281–306.

JUKES, T. H., 1987 Transitions, transversions, and the molecular evolutionary clock. J. Mol. Evol. **26:** 87–98.

JUKES, T. H., and V. BHUSHAN, 1986 Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J. Mol. Evol. **24:** 39–44.

KEMP, T. S., 1988 Interrelationships of the Synapsida, pp. 1–22 in *The Phylogeny and Classification of the Tetrapods*, Vol. II, edited by M. J. BENTON. Clarendon Press, Oxford.

KLUGE, A. G., and J. S. FARRIS, 1969 Quantitative phyletics and the evolution of anurans. Syst. Zool. **18:** 1–32.

LAKE, J. A., 1988 Origin of the eucaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature **331:** 184–186.

LAKE, J. A., 1989 Origin of the eukaryotic nucleus determined by rate-invariant analyses of ribosomal RNA genes, pp. 87–101 in *The Hierarchy of Life*, edited by B. FERNHOLM, K. BREMER and H. JORNVALL. Elsevier, Amsterdam.

LAWRENCE, C. B., and D. A. GOLDMAN, 1988 Definition and identification of homology domains. Comp. Appl. Biosci. **4:** 25–33.

LEWIN, B., 1987 *Genes III*. Wiley & Sons, New York.

LI, W.-H., C.-I WU and C.-C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

MANTEL, N. A., 1967 The detection of disease clustering and a generalized regression approach. Cancer Res. **27:** 209–220.

MARSHALL, C. R., 1992 Substitution bias, weighted parsimony, and amniote phylogeny as inferred from 18s rRNA sequences. Mol. Biol. Evol. **9:** 370–373.

MICKEVICH, M. F., and S. J. WELLER, 1990 Evolutionary character analysis: tracing character change on a cladogram. Cladistics **6:** 137–176.

MILNER, A. R., 1988 The relationships and origin of living amphibians, pp. 59–102 in *The Phylogeny and Classification of the Tetrapods*, Vol. I, edited by M. J. BENTON. Clarendon Press, Oxford.

MINDELL, D. P., and R. L. HONEYCUTT, 1990 Ribosomal RNA in vertebrates: evolution and phylogenetic implications. Annu. Rev. Ecol. Syst. **21:** 541–566.

MISHLER B. D., K. BREMER, C. J. HUMPHRIES and S. P. CHURCHILL, 1988 The use of nucleic acid sequence data in phylogenetic reconstruction. Taxon **37:** 391–395.

Molecular Biology Information Resource, 1989 *Eugene User's Manual, Release 3.2.* Department of Cell Biology, Baylor Medical College, Houston.

NELLES, L., B. L. FANG, G. VOLCKAERT, A. VANDENBERGHE and R. DEWACHTER, 1984 Nucleotide sequence of a crustacean 18S ribosomal RNA gene and secondary structure of eukaryotic small subunit ribosomal RNAs. Nucleic Acids Res. **12:** 8749–8768.

NOVACEK, M. J., A. R. WYSS and M. C. MCKENNA, 1988 The major groups of eutherian mammals, pp. 31–72 in *The Phylogeny and Classification of the Tetrapods*, Vol. II, edited by M. J. BENTON. Clarendon Press, Oxford.

OHTA, T., 1983 On the evolution of multigene families. Theor. Popul. Biol. **23:** 216–240.

OLSEN G. J., and C. R. WOESE, 1989 A brief note concerning archaebacterial phylogeny. Can. J. Microbiol. **35:** 119–123.

PATTERSON, C., 1989 Phylogenetic relations of major groups: conclusions and prospects, pp. 471–488 in *The Hierarchy of Life*, edited by B. FERNHOLM, K. BREMER and H. JORNVALL. Elsevier, Amsterdam.

RAYNAL, F., B. MICHOT and J.-P. BACHELLERIE, 1984 Complete nucleotide sequence of mouse 18S rRNA gene: comparison with other available homologs. FEBS Lett. **167:** 263–268.

SALIM, M., and B. E. H. MADEN, 1981 Nucleotide sequence of *Xenopus laevis* 18S ribosomal RNA inferred from gene sequence. Nature **256:** 205–208.

SMOUSE, P. E., J. C. LONG and R. R. SOKAL, 1986 Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool. **35:** 627–632.

SOGIN, M. L., and H. J. ELWOOD, 1986 Primary structure of the *Paramecium tetraurelia* small-subunit rRNA coding region: phylogenetic relationships within the Ciliophora. J. Mol. Evol. **23:** 53–60.

SWOFFORD, D. L., 1989 *PAUP: Phylogenetic Analysis Using Parsimony.* Illinois Natural History Survey, Champaign, Ill.

TEMPLETON, A. R., 1983 Convergent evolution and nonparametric inferences from restriction data and DNA sequences, pp. 151–179 in *Statistical Analysis of DNA Sequence Data*, edited by B. S. WEIR. Marcel Dekker, New York.

TORCZYNSKI, R. M., M. FUKE and A. P. BOLLON, 1985 Cloning and sequencing of a human 18S ribosomal RNA gene. DNA **4:** 283–291.

TURNER, D. H., N. SUGIMOTO and S. M. FREIER, 1988 RNA

structure prediction. Annu. Rev. Biophys. Biophys. Chem. **17:** 167–192.

VAWTER, LISA, 1991 Evolution of blattoid insects and of the small subunit ribosomal RNA gene. Ph.D. Thesis, University of Michigan, Ann Arbor.

WADA, A., A. SUYAMA and R. HANAI, 1991 Phenomenological theory of GC/AT pressure on DNA base composition. J. Mol. Evol. **32:** 374–378.

WILLMER, P., 1990 *Invertebrate Relationships: Patterns in Animal Evolution*. Cambridge University Press. Cambridge.

WOESE, C. R., 1989 Archaebacteria and the nature of their evolution, pp. 119–130 in *The Hierarchy of Life*, edited by B. FERNHOLM, K. BREMER and H. JORNVALL. Elsevier, Amsterdam.

WOESE, C. R., R. GUTELL, R. GUPTA and H. F. NOLLER, 1983 Detailed analysis of the high order structure of 16S-like ribosomal ribonucleic acids. Microbiol. Rev. **47:** 621–669.

ZUKER, M., 1989 Computer prediction of RNA structure. Methods Enzymol. **180:** 262–288.

Communicating editor: M. SLATKIN

APPENDIX A

The structural composition of the srDNA data set is given in Table 3.

TABLE 3

Structural composition of the srDNA data set

	Composition (%)				
	Stem	Loop	Bulge	Other	Unknown
HTU 7	27.5	6.1	9.4	17.1	39.9
HTU 8	44.0	9.7	14.6	18.3	13.4
HTU 9	44.9	10.7	17.0	19.7	7.7
HTU 10	45.4	10.7	16.9	19.7	7.2
C. elegans	42.2	8.5	13.6	22.1	13.7
A. salina	43.1	10.5	14.4	22.2	9.8
X. laevis	44.9	10.3	15.3	20.1	9.5
H. sapiens	44.6	10.7	18.0	19.6	7.1
R. rattus	45.3	10.7	14.5	22.6	7.0
M. musculus	45.0	10.7	14.2	23.0	7.2
OTUs	44.2	10.2	15.0	21.6	9.0
HTUs and OTUs	42.7	9.9	14.8	20.4	12.3

Structural composition of the entire data set, including all HTUs and unknown regions. As explained in the text, the large genetic distance between Caenorhabditis and Artemia accounts for the large percentage of positions belonging to the unknown category in HTU 7.