Paired Sequence Difference in Ribosomal RNAs: Evolutionary and Phylogenetic Implications[1]

Ward C. Wheeler and Rodney L. Honeycutt

Department of Organismic and Evolutionary Biology and the Museum of Comparative Zoology, Harvard University

Ribosomal RNAs have secondary structures that are maintained by internal Watson-Crick pairing. Through analysis of chordate, arthropod, and plant 5S ribosomal RNA sequences, we show that Darwinian selection operates on these nucleotide sequences to maintain functionally important secondary structure. Insect phylogenies based on nucleotide positions involved in pairing and the production of secondary structure are incongruent with those constructed on the basis of positions that are not. Furthermore, phylogeny reconstruction using these nonpairing bases is concordant with other, morphological data.

Introduction

The neutral hypothesis, as stated by Kimura (1968, 1969, 1983), suggests that most changes in nucleotide sequence occur in the absence of positive selection whereas those that are deleterious are removed by negative or "purifying" selection. The implications of such a theory are broad not only with respect to the interpretation of molecular evolutionary processes but also in the reconstruction of evolutionary patterns as reflected in phylogenetic trees.

RNA molecules provide a unique opportunity to examine the patterns of nucleotide sequence change. A large body of nucleotide sequences representing diverse organisms has been compiled (Erdmann et al. 1985) and small ribosomal RNA (rRNA) molecules have been used extensively in phylogenetic studies (Fox et al. 1980; Ohama et al. 1984; De Wachter et al. 1985). In addition, the structure of these molecules is such that one may address both the neutrality of nucleotide changes in RNAs and the effects that secondary structure might have on the derivation of phylogenies by means of sequence data. rRNAs form secondary structures through Watson-Crick pairing in helical duplex regions in the molecule, whereas the single-stranded or loop regions (fig. 1) are important for molecular recognition (Lewin 1983). Similar secondary structures occur over broad taxonomic groups (Erdmann et al. 1985); yet the ability to maintain these structures depends on the fidelity of pairing between distantly located bases, implying that any change involving paired bases is potentially deleterious. Using tRNAs as an example, both Ohta (1973) and Kimura (1983) have suggested that even though such changes may be slightly deleterious, they can be explained within the confines of the neutral theory. Such an explanation requires that two independent nucleotide changes involving paired bases occur at a frequency predicted by chance. If this is the case, it suggests a test of the neutral hypothesis.

1. Key words: ribosomal RNA genes, paired sequence difference, secondary structure, phylogeny reconstruction.

Address for correspondence and reprints: Ward C. Wheeler, Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138.

Mol. Biol. Evol. 5(1):90–96. 1988.

0737-4038/88/0501-0007$02.00

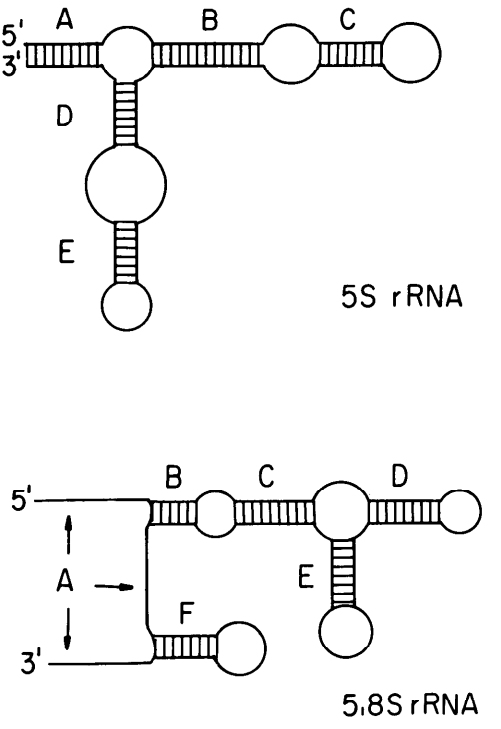

FIG. 1.—Schematic representations of the secondary structure of the 5S rRNA and 5.8S rRNA. The lettered areas, A–E in the 5S and A–F in the 5.8S, denote areas of double strand formation. Area A of the 5.8S rRNA base pairs with the 26S rRNA, whereas all the other areas form internal helices. The areas shown as "loops" between the double-stranded "stems" are single-stranded regions.

Here we show that there is a significant overoccurrence of sequence differences that maintain base pairing, suggesting that there is positive selection for the second, complementarity-restoring change. We have also investigated how these changes in paired regions affect phylogeny reconstruction compared with changes in single-stranded regions. Through this analysis, we determined (1) that the most parsimonious trees constructed on the basis of the double-stranded-region data can disagree with those constructed on the basis of the single-stranded-region data and (2) that the single-stranded-region phylogenetic trees are more concordant with trees based on other data.

Methods

To examine the distribution of paired and unpaired sequence differences in rRNAs and their effect on phylogeny reconstruction, we examined 5S and 5.8S rRNA sequences. Aligned sequences were from Erdmann et al. (1985), and D. L. Swofford's PAUP program (version 2.4) was used to examine the distribution of sequence differences by means of reconstruction of phylogenetic trees on the basis of a parsimony analysis. Parsimony involves the construction of phylogenetic trees of minimum length, in which length is the number of evolutionary events required to explain the current distribution of sequence differences among taxa.

Using 5S rRNA, we first attempted, through examination of well-established chordate (Wiley 1979), arthropod (Hennig 1969, 1981; Kristensen 1975, 1981; Boud-

reaux 1979), and plant (Crane 1985) phylogenies (fig. 2), to show the nonrandom occurrence of sequence differences that maintain base pairing. These preexisting trees were used as frameworks for reconstructing sequence differences. For each of these three cases, the most parsimonious scheme of sequence evolution was reconstructed on the basis of the sequence data including hypothetical ancestors. These ancestors are the nodes (branching points) of the trees. Through examination of the changes that occurred between nodes on these trees, we observed the behavior of paired sequence differences.

The crux of the test is that, for all cases in which both of the two base-paired nucleotides change, if the first change is neutral, then the second also should be so. If this is true, only one-third of the second changes should restore base pairing. (It is not necessary to know which change came first, because the second would still have to recreate what the first had destroyed.) In cases in which a paired couplet underwent a change in more than one path between tree nodes, only one path was counted. The difference between expected and observed distributions was compared for statistical significance via the χ^2-test. Although some of the expected cell sizes were small (<5), the G-test (Sokal and Rohlf 1981) was not used because several of the observed values were zero, thereby rendering the statistic incalculable.

The second type of analysis was performed on arthropods by using a different molecule, the 5.8S rRNA. Variable base positions were treated as unordered phylogenetic characters with five states: A, C, U, G, or 0 (= absent). Shorter trees, those

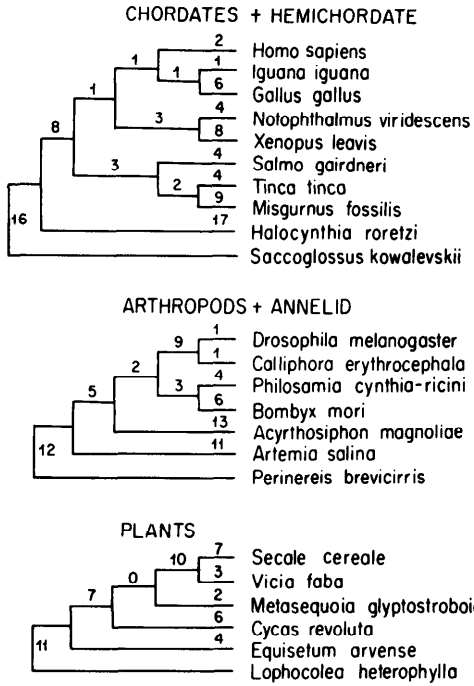

FIG. 2.—Chordate + hemichordate, arthropod + annelid, and plant phylogenies used to examine the distribution of sequence differences. The chordate + hemichordate phylogeny is that of Wiley (1979); the arthropod + annelid phylogeny is from Hennig (1969, 1981), Boudreaux (1979), and Kristensen (1975, 1981); and the plant phylogeny is from Crane (1985). The total number of sequence changes postulated to have occurred in each interval is noted above the path.

involving the lowest number of changes, were judged to be superior (more parsimonious) explanations of the data than the longer ones. In the analysis, the most parsimonious cladogram, as found by PAUP, was determined for three data sets. These data sets were (1) all the variable positions, (2) only those variable positions that base pair, and (3) only those variable positions that occur in single-stranded regions. In this way the relative effects of changes in single-stranded and duplex regions could be separated and compared with the overall effects of the entire data set on phylogenetic reconstruction. In making this distinction, we lumped all bases that form base pairs, both those that pair within the 5.8S molecule and those that form Watson-Crick pairs with the 26S rRNA. Three cladograms were produced; each was the most parsimonious for at least one of the data sets.

Results and Discussion

As shown in table 1, analysis of the chordate, arthropod, and plant 5S rRNA sequences demonstrates that there is a significant ($P < 0.001$) excess of base-paired differences in observed distributions over that expected by chance. In all three cases, the observed number of paired differences was three to four times that predicted by the neutral model. Concomitantly, the occurrence of unpaired differences was very rare.

The nonrandom pattern of paired differences in these rRNA sequences suggests that positive selection is driving the fixation of changes that restore base complementarity. We have shown this to be the case in the chordates/hemichordate, arthropods/annelid, and plants data sets. Presumably, this is a general property of RNAs having secondary structure formed through base pairing. Everything from the small tRNAs to much larger 18S and 26S rRNAs exhibits analogous secondary structures that rely on base pairing between nucleotide positions (Lewin 1983). Most likely, these molecules will also show a similar mode of evolution, in which one substitution is fixed and another complementary substitution is positively selected to ameliorate the negative effects of the first.

Table 1
Observed and Expected Substitutions in 5S rRNA

Data Set and Substitution	Differences		χ^{2a}
	Paired	Unpaired	
Chordates + hemichordate:			
Observed	12	2	15.2
Expected	5	9	
Arthropods + Annelid:			
Observed	12	0	24.0
Expected	4	8	
Plants:			
Observed	7	0	17.5
Expected	2	5	

NOTE.—Summary of observed vs. expected distributions of paired sequence differences (those that recreate base pairing) and unpaired sequence differences (those that are not base paired). These are the totals for all cases in which both positions that base pair have changed between adjacent nodes of the respective phylogenetic trees shown in fig. 2.

[a] 1 df; significant at $P < 0.001$.

Phylogenetic analyses of the 5.8S rRNA arthropod data show that the most parsimonious cladograms for single-stranded positions, double-stranded positions, and the complete data set are not the same (fig. 3). The minimum tree length for the single-stranded-position data was four to five steps shorter than that supported by either the double-stranded-position data or the complete data set.

Given that these RNA nucleotide positions are varying in tandem, it is not surprising to find that phylogenetic reconstruction is sensitive to the locations of the positions that are used to derive cladograms. Since base positions involved in pairing comprise two-thirds of the 5S rRNA molecule, the overall pattern of the data conforms to that of data on the double-stranded areas.

The relationships derived from data on the double-stranded regions of the 5.8S rRNA are an example of the molecular trees being incongruent with other data sets (fig. 3). Nearly all insect systematists support the grouping of figure 3C, derived from the loop data (Hennig 1969, 1981; Kristensen 1975, 1981; Boudreaux 1979). Of the insect taxa, *Drosophila* and *Sciara* are flies, *Bombyx* and *Philosamia* are moths, and *Acyrthosiphon* is an aphid. Both the complete and the stem-region data of the 5.8S rRNA place the brine shrimp, *Artemia,* and the aphid, *Acyrthosiphon,* with the Lepidoptera. In doing this, these data suggest breaking up seven well-established groups: Hexapoda, Insecta (= Ectognatha), Dicondylia, Pterygota, Neoptera, Holometabola, and the Panorpida. Members of the hexapodan taxa, unlike those of *Artemia,* have six legs. The five insects have external mouthparts and are dicondylic in that they

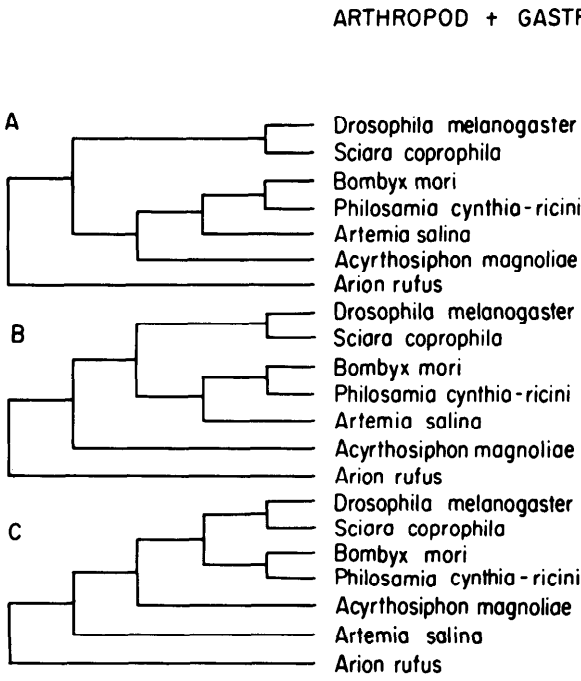

FIG. 3.—Three topologies found for the arthropod and gastropod 5.85 ribosomal RNA sequence data. Phylogeny A is the most parsimonious for the overall data. Phylogenies A and B are the most parsimonious for the sequence data from the double-stranded regions of the molecule, whereas the most parsimonious reconstruction of sequence difference in the single-stranded nucleotide positions is shown in phylogeny C. The overall number of sequence differences required by each topology is shown for all bases, stem (double-stranded) bases, and loop (single-stranded) bases.

have two mandibulary articulation points. Again unlike the brine shrimp, they have wings and these wings fold over the abdomen. Furthermore, the flies and moths share complete development, the origin of wings in larval invaginations, and several mouthpart characters. The possibility that these characters have originated through convergence is remote.

The patterns in the phylogenetic analyses of the group of organisms examined in the present study are clear—double-stranded and single-stranded sequence changes (or characters) led to different, if not misleading, phylogenetic conclusions. Furthermore, since these nucleotide data could be compared with independent data sets, single-stranded nucleotide positions seem to be a more accurate reflector of genealogy. When the frequency of pairing in differences is high, the disinformation content of the data set becomes great enough to result in misleading phylogenetic hypotheses. In a study of tRNA evolutionary behavior, Holmquist et al. (1973) found no consistent patterns of relationship among the sequences that they examined. Although they distinguish between helical and nonhelical nucleotide positions, no connection is postulated between higher-order structure and the lack of resolution in their phylogenetic analysis. This may be occurring because the interdependency of differences destabilizes the data set. When characters are functionally linked or part of some ontogenetic complex, they are only expressing a single source of information. Only independently varying qualities are viable characters in phylogenetic reconstruction. In the case of paired differences, the maleffects are amplified by an additional factor: there are two to three times as many potentially varying positions in the double-stranded areas as in the single-stranded regions. This leaves the data set biased in favor of the positions that pair. The independent data are overwhelmed by these coevolving bases.

As a result of such character covariance, we should suspect any cladogram based on RNA sequence data that does not consider this problem. Such a situation may have occurred in molecular phylogenies of many groups. Other, larger, rRNAs (e.g., the 16/18S) also have involved secondary structures and presumably exhibit the effects that we have seen in our analysis. Therefore, if we are correct and if the patterns that we see are general, schemes of relationship based on this type of sequence information may require revision.

To avoid misleading phylogenies based on paired differences, nucleotide positions that pair should be downweighted, perhaps by one-half, or even excluded entirely. This may seem extreme, but since these positions can be disinformative owing to their sensitivity to selection pressures, there is no other recourse.

The popularity of using rRNAs in systematics is increasing. New data will allow us to test the generality of these observations and to determine both whether paired differences occur as a matter of course and whether they impede systematic inference. The analyses presented herein should aid in the examination of molecular data and increase the confidence that we place in our phylogenies.

Acknowledgments

We would like to thank T. F. Smith, M. Ruvolo, A. H. Knoll, S. W. Schaeffer, R. J. Baker, J. A. Birchler, J. O. Carpenter, K. Nelson, A. J. Werth, J. S. Jensen, E. L. Broach, W. M. Fitch, and two anonymous reviewers for comments and criticisms on the manuscript. This research was supported by National Science Foundation grant BSR 85-08479 to R.L.H. and by National Science Foundation Graduate Fellowship and grant BSR 87-00966 to W.C.W.

LITERATURE CITED

BOUDREAUX, H. B. 1979. Arthropod phylogeny with special reference to insects. John Wiley & Sons, New York.

CRANE, P. R. 1985. Phylogenetic relationships in seed plants. Cladistics **1**:329–348.

DE WACHTER, R. E. HUYSMANS, and A. VAN DEN BERGHE. 1985. 5S Ribosomal RNA as a tool for studying evolution. Pp. 115–141 *in* K. H. SCHLEIFER and E. STACKEBRANDT, eds. Evolution of prokaryotes. Academic Press, London.

ERDMANN, V. A., J. WOLTERS, E. HUYSMANS, and R. DE WACHTER. 1985. Collection of published 5S, 5.8S, and 4.5S ribosomal RNA sequences. Nucleic Acids Res. **13**(Suppl.): r105–r153.

FOX, G. E., E. STACKEBRANDT, R. B. HESPELL, J. GIBSON, J. MANILOFF, T. A. DYER, R. S. WOLFE, W. E. BALCH, R. S. TANNER, T. J. MAGRUM, L. B. ZABLEN, R. BLAKEMORE, R. GUPTA, L. BONEN, B. J. LEWIS, D. A. STAHL, K. R. LUEHRSEN, K. N. CHEN, and C. R. WOESE. 1980. The phylogeny of the prokaryotes. Science **209**:457–463.

HENNIG, W. 1969. Die Stammesgeschichte der Insekten. Seckenberg-Büchern. **49**:1–436.

———. 1981. Insect phylogeny. John Wiley & Sons, New York.

HOLMQUIST, R., T. H. JUKES, and S. PANGBURN. 1973. Evolution of transfer RNA. J. Mol. Biol. **78**:91–116.

KIMURA, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624–626.

———. 1969. The rate of molecular evolution considered from the standpoint of population genetics. Proc. Natl. Acad. Sci. USA **63**:1181–1188.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, New York.

KRISTENSEN, N. P. 1975. The phylogeny of hexapod "orders": a critical review of recent accounts. Z. Zool. Syst. Evol. Forsch. **13**:1–44.

———. 1981. Phylogeny of insect orders. Annu. Rev. Entomol. **26**:135–157.

LEWIN, R. 1983. Genes. John Wiley & Sons, New York.

OHAMA, T., T. KUMAZAKI, H. HORI, and S. OSAWA. 1984. Evolution of multicellular animals as deduced from 5S rRNA sequences: a possible early emergence of the metazoa. Nucleic Acids Res. **12**:5101–5108.

OHTA, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature **246**:96–98.

SOKAL, R. R., and F. J. ROHLF. 1981. Biometry. W. H. Freeman, New York.

WILEY, E. O. 1979. Ventral gill arch muscles and the interrelationships of the gnathostomes, with a new classification of the Vertebrata. J. Linnaean Soc. **67**:149–180.

WALTER M. FITCH, reviewing editor

Received September 29, 1987