An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA

Daniel L.Nickrent* and Malcolm L.Sargent
Department of Plant Biology, University of Illinois, Urbana, IL 61801, USA

Received August 9, 1990; Revised and Accepted November 28, 1990

ABSTRACT

The V4 region of the small subunit (18S) ribosomal RNA was examined in 72 different sequences representing a broad sample eukaryotic diversity. This domain is the most variable region of the 18S rRNA molecule and ranges in length from ca. 230 to over 500 bases. Based upon comparative analysis, secondary structural models were constructed for all sequences and the resulting generalized model shows that most organisms possess seven helices for this region. The protists and two insects show from one to as many as four helices in addition to the above seven. In this report, we summarize secondary structure information presented elsewhere for the V4 region, describe the general features for helical and apical regions, and identify signature sequences useful in helix identification. Our model generally agrees with other current concepts; however, we propose modifications or alternative structures for the start of the V4 region, the large protist inserts, and the sector that may possibly contain a pseudoknot.

INTRODUCTION

The primary and higher order structures of the small-subunit ribosomal RNAs have been determined for a large number of eukaryotic organisms (see 1, 2 for recent compilations). With database archives as well as unpublished sequences, over 200 eukaryotic 18S rRNA sequences have been determined. Despite this large number of sequences, elucidation of secondary structural features in regions displaying high variability has not been straightforward. The work of Woese et al. (3) first established the basic folding structures the prokaryotic 16S rRNA and Brimacombe (4) showed that the eukaryotic 18S rRNA molecule could be folded into a similar configuration. The region termed V4, following the nomenclature of Nelles et al. (5) and subsequent papers from the laboratory of R. DeWachter, is the largest and most complex of the highly variable regions. The region is present in most prokaryotic and organellar small-subunit rRNAs as a single helix of ca. 60 bases. In eukaryotes, the length of the V4 region ranges from ca. 230 bases (the most common state) to ca. 520 bases in several protists. Exceptions include two parasitic flagellates: *Giardia lamblia* (6), which has an extremely reduced V4 region and *Vairimorpha necatrix* (7) which lacks the V4 region entirely.

Determination of the secondary structural features for the V4 region has been hampered mainly because sequences from the wide array of organisms required for a comparative approach (8,9) have been lacking. The first helical model proposed for the V4 domain was from *Saccharomyces* and *Xenopus* by Zwieb et al.(10). Subsequently, a model for the nuclear 18S rRNA of *Artemia*.was proposed (5) giving three helices in this region and leaving two areas undefined. During the same year, several secondary structural features of the 18S rRNA of *Xenopus* were presented (11) and one helix in the V4 region (21-C) was confirmed by ribonuclease analysis. *A* complete secondary structure for the V4 region of the nemotode *Caenorhabditis elegans* (12) agreed with that of Nelles et al (5) for *Artemia* and included three additional helices in the areas left undefined by them. Gonzalez and Schmickel (13) put forth a model for the human 18S rRNA that differed in many respects from those proposed for *Xenopus* and *Caenorhabditis*.

The compilation of 15 eukaryotic small-subunit rRNA sequences reported by Huysmans and De Wachter (14) included a schematic diagram for three helices in the V4 region based on the work of Nelles et al. (5). This alignment was expanded to 40 sequences by Dams et al. (1) who included a diagramatic model of the V4 region incorporating four additional helices for the two areas previously left undefined. Two of these helices are common to all eukaryotic organisms and the remaining two are found only in those protists with large insertions. The most recent compilation of 62 sequences (2) further modifies the above model to include one additional helix (E21-2) and a pseudoknot incorporating elements of two adjacent helices.

In the present study, we present an overview of the structure of the V4 domain, indicate signature features that help characterize helical and loop regions, and propose some modifications of and alternatives to the secondary structure of this region as reported by Neefs et al. (2). We have used a comparative approach to evaluate helical and nonhelical regions from sequences of eukaryotes that span the diversity of taxonomic groups. A recent paper by Neefs and DeWachter (15) also

* To whom correspondence should be addressed at Department of Plant Biology, Southern Illinois University, Carbondale, IL 62901, USA

provides an overview of the 5–6 helices of this V4 region that are found in almost all eukaryotes. Their extensive, computational analysis compliments the more taxonomic approach that we have employed and increases the confidence level with which several of these helices may be approached. Our conclusions are similar to theirs in many regards; the differences will be discussed in detail below.

This examination of the V4 region was undertaken to assist in identifying signatures in secondary structure and generating valid alignments useful in phylogenetic studies. Most phylogenetic analyses of molecular data utilize only the primary sequence as an input matrix for analysis. However, the fact that base changes in helical regions are not independent can be used to weight characters (16–19). Knowledge of secondary structural features is also critical in allowing advancement in understanding of covariances and tertiary interactions (20,21) as well as ribosome function (22).

MATERIALS AND METHODS

We have utilized all of the sequences in Neefs et al. (2), additional sequences available from GENBANK or EMBL, as well as unpublished sequences determined by our laboratory and others (Table 1). The initial sequence alignments were obtained from R. Gutell and subsequent modifications were performed using a Sun workstation and an alignment program written by Thomas Macke. Further sequence manipulation was carried out on an IBM-AT using the Eyeball Sequence Editor of Eric Cabot. Our alignment for this region is available upon request from the authors. Secondary structural diagrams were composed for all 72 sequences (Table 1) using a Mac IIcx and MacDraw software. Not all secondary structures are reproduced here, but ones corresponding to published sequences may be obtained upon request.

When referring to base pair sites, numbering begins with the most 5′ base of the helix and both symmetrical and asymmetrical unpaired bases are given site numbers. We realize that maximum base pairing, e.g. as determined from free energy analysis (30), does not always reflect the true structure of the rRNA in the ribosome, but in the absence of additional information, we have opted for maximum pairing. Both canonical and noncanonical (G•U and G□A) pairing was allowed. For four helices (21a, 21b, 21-6, and 21-7; see Figure 1) we have constructed matrices comparing base pairing for each taxon at each site. These matrices are available from the authors upon request.

Our approach in dealing with the extensive noise, i.e. the very large number of insertions, deletions, and substitutions, in this region has been, in the absence of other information, to search for helical configurations in which compensating base changes are group specific. We have assumed that the major taxonomic groups delimited in Table 1 are valid on the basis of other criteria and that group-specific changes are informative and a means to deal with the noise. This approach is thus different from, but complimentary to, the approach of Neefs and DeWachter (in press, this journal) who have done a computational search for those helical configurations that maximize base pairing from all the models that have been proposed to date. Transposibility, the ability to fold the molecule from all taxa according to the consensus model, is also considered by us, as well as Neefs and De Wachter, to be an important criterion in determining the validity of a particular helix. There may be exceptions, however, to the use of this criterion and they are discussed below.

RESULTS AND DISCUSSION

A generalized model for the eukaryotic V4 region, based on Neefs et al. (2), includes a numbering system for all helices identified to date (Figure 1). Features of this model will be discussed in turn beginning with the 5′ end. We have accomodated the addition of helices not reported in Neefs et al. (2) by adding a letter suffix, e.g. 21-1a. The overall configuration of the 5–6 major helices found in almost all eukaryotic organisms is, of all the models surveyed by Neefs and De Wachter (15), most like that proposed by Ellis et al. for *Caenorhabditis* (12) although there are some significant, specific differences in helices 21, 21-6 and 21-7.

Helix 21

Our model for this helix (Figure 2a; the human sequence, but identical to the eukaryotic consensus sequence) is based upon Gutell (personal communication). The model presented in figures 17–21 in Gutell et al. (9) differs from the above in that errors in the primary sequence are present in several taxa. Figure 2a differs from the model proposed by Neefs et al. (2) in Figure 2b in two respects. First, helix 21 is shown starting on the fifth of five A's in the 5′ strand vs. a U (a four base slide in the 3′ direction). Second, we propose that the helix extends several bases beyond the junction with helix 22 and define this extension as 21b. The model of Neefs et al. (2) is similar in configuration to that of the prokaryotic molecule and is thus supported in that the prokaryotic secondary structure models have, in general, been good predictors for their eukaryotic counterparts. It should be emphasized, however, that this eukaryotic V4 region is extremely different from the single helix found in prokaryotes, so it is quite conceivable that the 'lead-in' region, i.e. helix 21, is also different. In general, helix 21 is extremely conserved and the few compensating changes found in multicellular organisms do not allow a clear choice to be made between the two models.

Our model is favored, however, when compensating changes in several of the unicellular protists (the amoeboid-like organisms and the flagellates) are considered. The following changes in helix

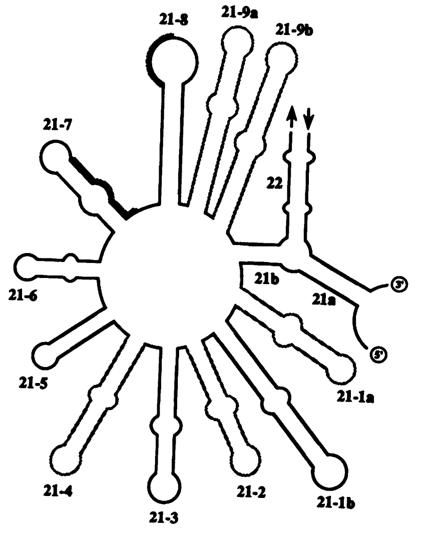

Figure 1. Generalized model for the eukaryotic V4 region of the the small-subunit rRNA. Helices found in all eukaryotes (*Giardia* excepted) are shown with bold lines whereas those found in a minority of taxa are indicated by shadowed lines. The helix numbering follows that of Neefs et al. (2). Shaded portions of helices 21-7 and 21-8 can be formed into a pseudoknot.

21a are noted: site 1 [A□G to C-G (*Drosophila*, *Euglena*, *Giardia*, *Physarum* and the 6 *Plasmodium* genes), to G-C (*Crithidia*, *Leishmania*, and *Trypanosoma*) or to U•G (*Euplotes*)]; site 2 [G-C to A-U (*Euglena*)]; site 3 [unpaired C, A to U-A (*Drosophila*, *Crithidia*, *Leishmania*, *Trypanosoma*, and *Plasmodium falciparum*, C gene) or to C-G (*Styela*, *Physarum*, and *Giardia*]; site 4 [U-A to C-G (*Naegleria* and *Giardia*)]; site 5 [C-G to U•G (*Euglena*)]; site 6 [G□A to G•U (*Euplotes*)]; sites 7–10 [conserved]; site 11 [U-A to C-G (*Tenebrio*, *Physarum*, and *Euglena*), to U•G (*Branchiostoma*, *Chlamydomonas*, *Volvox*, *Prorocentrum*), or to A-U (*Naegleria*)]. It should also be noted that this helix is extended by a G-C base pair in *Crithidia*, *Leishmania*, and *Trypanosoma*. This helix in *Giardia* appears to be extended by three base pairs.

Compensating or group-specific changes in helix 21b include: site 1 [G•U to A-U (21 taxa, primarily fungi, ciliates, and the amoeboid-like organisms)]; site 2 [no changes]; site 3 [U-A to unpaired C (A), A (20 taxa including 5 of 7 vascular plants, the two ascomycetes, all *Tetrahymena*, *Crithidia*, *Leishmania*, and *Trypanosoma*), or to U•G (*Caenorhabditis*, *Artemia* and *Euplotes*)]; site 4 [C-G to U•G (30 taxa including all eight algae, 4 of 5 fungi, all ciliates, and 5 of 9 of the amoeboid-like organisms), to A□G (*Artemia* and *Plasmodium falciparum*) or to U-A (Plasmodium berghei C gene)]; site 5 [no compensating changes]; site 6 [conserved]; site 7 [unpaired in 39 taxa to U-A (14 taxa including all vascular plants and the two ascomycetes), or to G□A (8 taxa, primarily in the invertebrates and flagellates)].

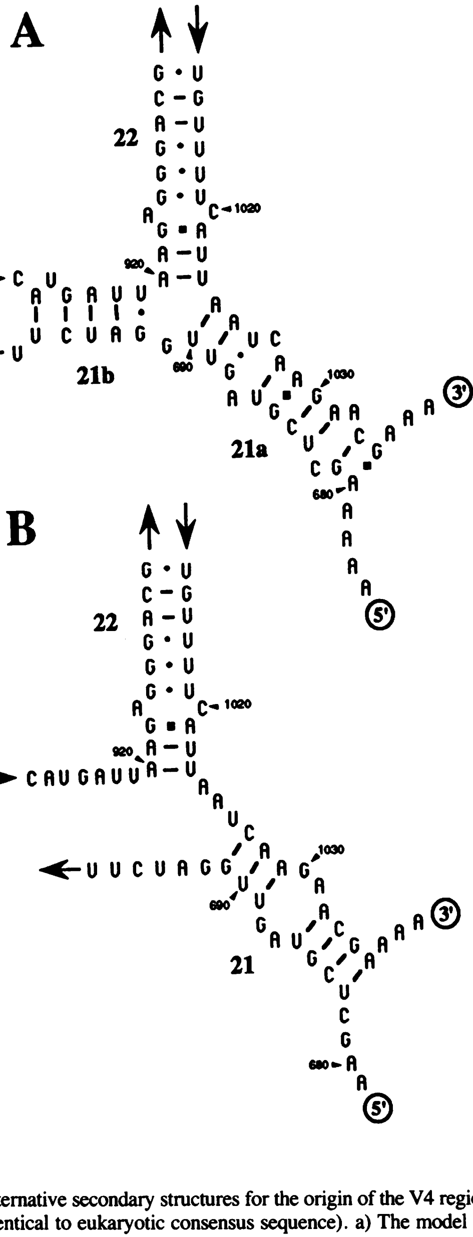

Figure 2. Alternative secondary structures for the origin of the V4 region (human sequence; identical to eukaryotic consensus sequence). a) The model favored in this paper showing helix 21 composed of two parts, 21a and 21b. b) The model for helix 21 as proposed by Dams et al. (1).

Helix 21b is very different in the flagellates. In *Crithidia*, *Leishmania* and *Trypanosoma*, there is an interstitial A as opposed to a G (as found in all other eukaryotes), sites 1 and 2 are missing, and site 3 is unpaired. In *Euglena* and *Naegleria*, there is apparently a short neck of three base pairs, different in the two species, which bears no obvious relationship to the consensus helix. In *Giardia*, this helix seems to be lacking entirely. In addition, the sequences on both sides of this helix in *Physarum* are very unlike those of all other eukaryotic organisms, but some pairing is possible (Figure 3).

We have considered several alternative configurations for the helix 21/ 22 junction across all taxa. It is possible to pair the 3′ bases of helix 21b with the 3′ bases of helix 21a at one base increments from base 919 (U) with base 1024 (A) to base 915 (U) with base 1028 (A) (Figure 2). In addition, stepwise pairing in the opposite direction is also possible from base 691 (G) with base 1023 (U) to 692 (G) with base 1022 (U). Given the high degree of conservation within these helices, it is not possible to determine the exact position where helix 21b emerges from 21a/22. On the basis of maximization of base pairing, however, the configuration shown in Figure 2a is optimum.

Two different interpretations have been proposed for the region including helices 20, 21, 22, 25, 26, 27 and 28 (30) that involve a possible structural switch. Our model (Figure 2a) does not provide evidence supporting one of these two structures over the other.

Helix 21-1 through 21-4

The sector spanning helices 21-1 through 21-4 (Figure 1) is the most variable in terms of the number of helices and primary sequence within the helices for the V4 domain. Helix 21-3 is the most conserved and was therefore identified first by Zwieb et al.(10) Subsequently, both Gonzalez and Schmickel (13) and Ellis et al. (12) proposed structures for the 21-1b helix in human and the nematode, respectively. A large number of taxa have many additional bases in this sector. Dams et al. (1) proposed helix 21-4 to accomodate these bases in several protists. Neefs et al. (2) added helix 21-2 for *Drosophila*.

For the vast majority of multicellular organisms and ciliates, analysis of this sector is straightforward and sequence alignments can be made with confidence. In these organisms, there are two helices (21-1b and 21-3). Of those helices universally present in the V4 region of eukaryotes, helix 21-1b is the most variable. The helix seldom exceeds 25 base pairs in length and often shows a G-rich area in the 5′ strand at the base of the helix. Unlike helices 21-3, -5, -7 and -8, the apex loop of this helix does not appear to have a distinct signature that applies to all eukaryotes. Within certain groups, however, signature sequences were seen, such as the CGU(G)YC in many unicellular organisms. Conserved base pairs in the stem of the helix can be seen only within certain groups, for example, in helix 21-1b

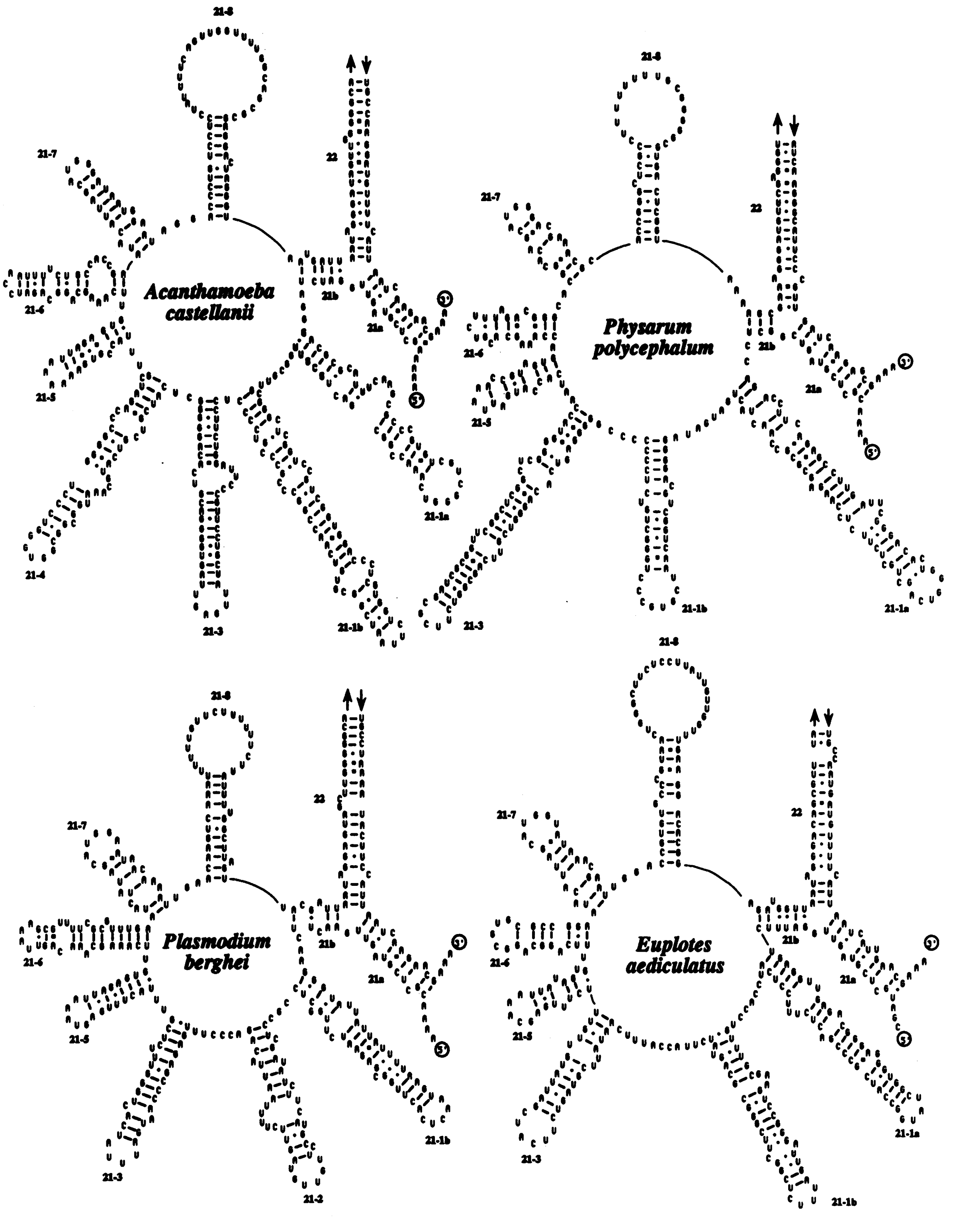

Figure 3. Secondary structural models for the V4 region of *Acanthamoeba castellanii*, *Physarum polycephalum*, *Plasmodium berghei*, and *Euplotes aediculatus*. All show additional helices relative to the more common state in eukaryotes between helices 21 and 21-5.

of multicellular organisms, the following base pairs are noted: sites 2 and 3 (G-C), site 15 (C-G), sites 17–20 (G-C, U-A, C-G, C-G, respectively). Helix 21-3 is generally ca. 20 base pairs in length and has an apical loop composed of ca. six bases. A conserved sequence in the apex loop is U,U, N, A (or U, U, N, G in a few taxa). Signature base pairs within the stem of this helix can be defined only within certain taxonomic groups. For example, in the multicellular organisms, the following conserved sites are noted: site 5 (G-C), site 6 (G-C), site 15 (U•G), and site 19 (C-G).

For those problem taxa (flagellates, amoeboid-like organisms, *Euplotes* and insects) with large inserts in this sector, we have developed models that involve three or four helices of lengths typical of the V4 region. Our strategy has been to search first

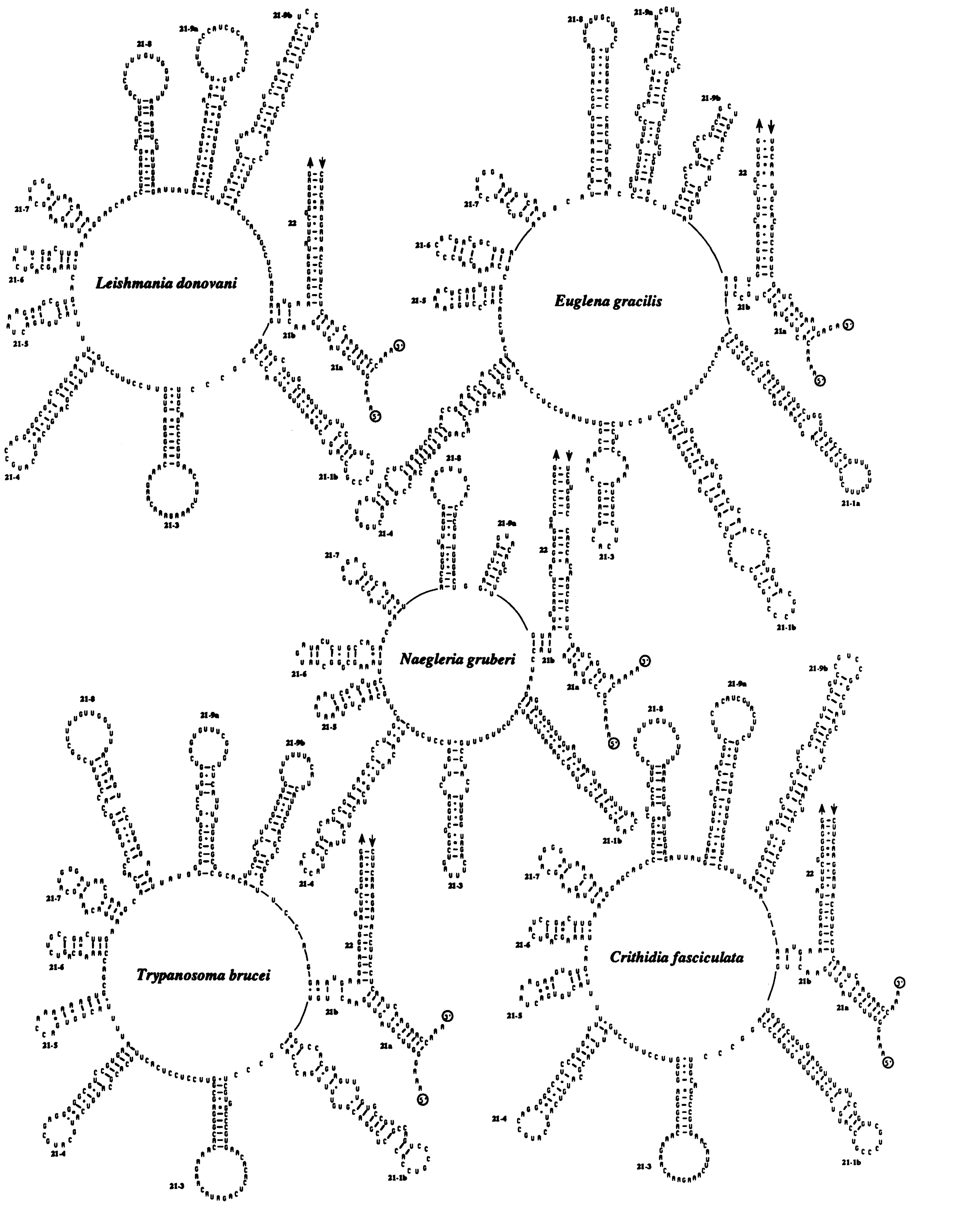

Figure 4. Secondary structural models for the V4 region of five flagellates showing additional helices (21-1a, 21-4, 21-9a, and 21-9b) in comparison with most other eukaryotes.

for helix 21-3 using the apical loop signature sequence noted above followed by identification of helix 21-1b. For *Physarum*, we found three helices identified as 21-1a, 21-1b, and 21-3 (Figure 3). Given the basal grouping of Gs in the second helix, we considered it to be 21-1b. The resulting bases 5′ to this helix can then be formed into a new helix (21-1a). Based on this line of reasoning, we propose that three other taxa (*Acanthamoeba*, *Euglena* and *Euplotes*, Figures 3 and 4) possess helix 21-1a. In the apex loop of helix 21-1a, *Acanthamoeba* and *Physarum* share the sequence GGGUCA.

Our model for *Physarum* contrasts with the published alignment (2) which has two helices, one of which incorporates the bases in our helices 21-1a and -1b. For *Acanthamoeba* and *Euglena*, our four-helix models contrast with the three-helix models of Dams et al. (1). Their helix 21-1 is essentially equivalent to our helix 21-1a and they show two very long helices (21-2 and -3, their numbering) that incorporate the bases we have distributed in helices 21-1b, -3, and -4. For *Euplotes*, both our model and that of Dams et al. (1) utilize three helices, but these helices differ in detail and number designation. The third helix contains the signature UUNA, hence we consider this helix 21-3. The first helix we designate 21-1a because of sequence similarity between it and *Euglena* at the distal end of the stem (GGGUGGC). The middle helix in *Euplotes* could be assigned to either 21-1b or 21-2 since it lacks Gs in the 5′ strand at the base of the helix. At present we favor helix 21-1b since the alternative would result in *Euplotes* being the only eukaryote lacking this helix.

Hendriks et al. (31) proposed one helix to accommodate the unusually large number of bases between helices 21 and 21-3 for the flour beetle *Tenebrio*. The sequences of the rRNA genes for a second insect (*Drosophila melanogaster*) have now been reported (33) and Neefs et al. (2) proposed two helices (21-1 and 21-2) for these bases. We agree with Neefs et al. (2) and suggest that a similar model can be applied to *Tenebrio* . There is clear evidence of homology between the 21-3 helices of the two species, but for helix 21-2 there is not. Strangely, apical (UUGUA) and stem features (UUUU bulge in the 3′ strand) of helix 21-2 are more similar for *Plasmodium berghei* and *Drosophila* than between the latter and *Tenebrio*. We assume that this is merely coincidental, although we note that only the insects and *Plasmodium* have this helix.

In *Plasmodium*, there is a great deal of variability between species and the A versus C genes in the area between helices 21 and 21-4. Neefs et al. (2) show an extremely long 21-1b helix (97 bases) with three major asymmetrical internal loops and a helix 21-3 of more typical length. As an alternative, we suggest that these available bases be placed in three helices (Figure 3). Probable assignments for these helices are 21-1b, 21-2, and 21-3 as determined by the 5′ basal Gs in *P. berghei*, the A, A, C, U in the apex loop of 21-1b in *P. berghei* and *P. falciparum*, and the UUNA of the apex loop of 21-3 of all three species. Despite the great interspecific variability, homology is very strong between *P. falciparum* and *P. lophurae* in the helical portion of 21-3 and the interstitial region 5′ to this helix.

Helix 21-4 is only present in the flagellates and *Acanthamoeba* (Figures 3 and 4), ranges from 14 to 21 base pairs in length (30 in *Euglena*), and has an apex loop ranging in size from five to nine bases. Homology is evident in this helix for *Crithidia*, *Leishmania*, and *Trypanosoma* but is not apparent for the other taxa. A signature of GC(A)UG exists in the apex loop of four of the five flagellates.

Helix 21-5

This is the most conserved helix in the V4 region among the eukaryotes and was first proposed by Nelles et al. (5). It is eight base pairs in length and characteristically possesses an A rich apex loop, typically with four or five consecutive As. The highly reduced parasite *Giardia lamblia* retains this and the three subsequent helices.

Helix 21-6 and 21-7

Aside from the areas of the V4 region that include the protist inserts, this region has remained the most perplexing in terms of secondary structure determination. Three different concepts for secondary structure in this area have been suggested: a) two helices [(12) based on *Caenorhabditis*]; b) one long helix [(31) based on*Tenebrio*]; or c) one short helix and a pseudoknot (2).

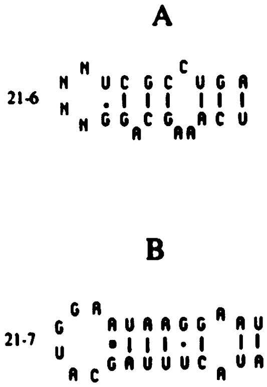

Figure 5. Consensus secondary structural models for a) helix 21-6 and b) 21-7 of the eukaryotic V4 region of the 18S rRNA.

We favor the concept of a single helix (21-6) for the beginning portion of this area, but remain neutral concerning acceptance of a pseudoknot versus a second helix. Our consensus model (Figure 5a) for helix 21-6 is similar in most respects with the one proposed for the human molecule by Gonzalez and Schmickel (13). In comparison with the model of Neefs et al. (2), ours shows fewer and smaller asymmetrical internal loops. In addition, we have examined both models and found that ours better explains the compensating changes seen across all taxa. The following compensating or group specific changes are noted: site 1 [U-A to U•G (13 taxa, generally in vascular plants and mammals) or to C-G (*Physarum*)]; site 2 [C-G to U•G (13 taxa, sporadically distributed)]; site 3 [A-U to C-G (four ciliates and *Giardia*) or to A□G (the birds, reptiles, and *Physarum*)]; site 4 [unpaired A to A-U (11 taxa including the four non-green algae, *Crithidia*, *Leishmania*, *Trypanosoma* and *Plasmodium berghei* A and C genes)]; site 5 [unpaired A, C to A-U (26 taxa, including all vascular plants, all green algae, all fungi, and all 6 *Plasmodium* genes) or to G-C (all ciliates)]; site 6 [G-C to G•U (6 taxa including *Ochromonas*, 4 of 6 *Plasmodium* genes, and *Giardia*), to G□A (*Physarum*, *Crithidia*, *Leishmania*, and *Trypanosoma*) or to A-U (*Plasmodium falciparum* C gene)]; site 7 [C-G to A□G (*Neurospora*)]; site 8 [no compensating changes]; site 9 [G-C to G•U (12 taxa, primarily in the fungi and protists), to G□A (8 taxa, sporadically distributed), to A-U (*Plasmodium berghei* A and C genes, *P. lophurae*, *Zamia* and *Caenorhabditis*) or to U-A (*Euglena*)]; and site 10 [G•U to G□A (15 taxa, primarily in the vascular plants, algae, and fungi) or to G-C (six taxa including all the mammals)].

Length variation is common in helix 21-6. The helix is shorter (missing site 10) in five different taxa and *Plasmodium vivax* shows an extreme condition in that it lacks sites 7–10. In contrast, the helix can be longer, e.g. it is extended by one C-G base pair at the terminus of the stem in mammals and *Xenopus*. In *Acanthamoeba*, *Naegleria*, and in five of the six *Plasmodium* genes, the extension involves 2 to 6 additional base pairs. Neefs et al. (2) have also used these additional bases by extending their helix 21-6, however, our model has fewer and smaller asymmetrical internal loops. In general, there are three bases in the apex loop of the protists, fungi, algae, and vascular plants, and four bases in the loop for the invertebrates and chordates.

Table 1. Small-subunit rRNA Sequences Examined

Taxon	Reference
Chordates	
Homo sapiens	2
Mus musculus	1
Oryctolagus cuniculus	2
Rattus norvegicus	1
Gallus gallus	23
Turdus migratorius	23
Alligator mississippiensis	23
Heterodon platyrhinos	23
Ambystoma mexicanum	23
Xenopus laevis	1
Latimeria chalumnae	24
Lepomis cyanellus	*
Petromyzon marinus	*
Branchiostoma floridae	*
Invertebrates	
Styela plicata	*
Asterias forbesi	25
Artemia salina	1
Eurypelma californica	2
Drosophila melanogaster	2
Tenebrio molitor	2
Lumbricus sp.	25
Spisula sp.	25
Golfingia gouldi	25
Lingula reevi	25
Riftia pachyptila	25
Caenorhabditis elegans	1
Dugesia tigrina	25
Hydra sp.	25
Fungi	
Neurospora crassa	1
Saccharomyces cerevisiae	1
Achlya bisexualis	1
Blastocladiella emersonii	25
Pneumocystis carinii	2
Vascular Plants	
Arabidopsis thaliana	2
Glycine max	1
Lycopersicon esculentum	2
Phoradendron serotinum	27
Oryza sativa	1
Zea mays	1
Zamia pumila	2
Algae	
Chlamydomonas reinhardtii	1
Chlorella vulgaris	2
Nanochlorum eukaryotes	2
Volvox carteri	2
Costeria costata	28
Ochromonas danica	1
Prorocentrum micans	29
Skeletonema costatum	2
Ciliates	
Euplotes aediculatus	1
Oxytricha nova	1
Paramecium tetraurelia	1
Stylonychia pustulata	1
Tetrahymena borealis	1
T. pigmentosa + 4	1**
T. pyriformis	1
T. thermophila + 1	1†
T. tropicalis	1
Amoeboid-like organisms	
Acanthamoeba castellanii	1
Dictyostelium discoideum	1
Physarum polycephalum	2
Plasmodium berghei/A gene	1
P. berghei/C gene	1
P. falciparum/A gene	2
P. falciparum/C gene	2
P. lophurae	2
P. vivax/A gene	30
Flagellates	
Crithidia fasciculata	1
Leishmania donovani	2
Trypanosoma brucei	1
Euglena gracilis	1
Giardia lamblia	2
Naegleria gruberi	2

* Stock and Whitt, unpublished

** *Tetrahymena hegewischi, T. australis, T. capricornis,* and *T. patula* V4 rRNA sequence identical to *T. pigmentosa*

† *Tetrahymena malaccensis* V4 rRNA sequence identical to *T. thermophila.*

The size of the apex loop varies somewhat due to occassional additional bases and because site 10 (and to a lesser extent site 9) of the stem often contains non-canonically paired or unpaired bases.

The existence of a distinct helix 21-7 as opposed to those bases being a component of a pseudoknot [labeled helix 21-7 by Neefs et al. (2)] remains uncertain, however, we are compelled to present the evidence supporting the former concept. First, primary and secondary structural features are conserved within the taxonomic groups shown in Table 1. Second, complementary changes occur in six of the nine sites within the helix. In our consensus model (Figure 5b), the helix is nine base pairs in length and the apex loop contains a conserved signature of C, A, U, G, G, A (C, U, A, G, G, A in mammals). The following compensating or group specific changes were found: sites 1 [A-U to A□G (*Crithidia, Leishmania, Trypanosoma,* and *Euglena*) or to G•U (*Naegleria*)]; site 2 [U-A to C-G (*Physarum* and *Naegleria*)]; site 3 [unpaired A, A to A-U (6 of 7 vascular plants, *Crithidia, Leishmania,* and *Trypanosoma*) or to A□G (all algae and *Euglena*)]; site 4 [C-G to unpaired C, A (18 taxa including 5 of 7 vascular plants, 6 of 8 algae, and 5 of 6 *Plasmodium* genes)]; site 5 [U•G to A-U (16 taxa including 4 of 5 fungi and 8 of 9 ciliates), to unpaired A, C (all vascular plants and green algae), to C-G (10 taxa including all mammals), to U-A (5 of 6 *Plasmodium* genes), or to G•U (*Ochromonas, Prorocentrum* and *Skeletonema*)]; site 6 [U-A to G•U (all mammals, 11 of 24 invertebrates and chordates)]; site 7 [U-A to unpaired C opposite A,A (13 of 14 chordates) or to G□A (9 of 14 invertebrates)];

site 8 [A-U to U-A (10 of 14 invertebrates)]; site 9 [G□A to G•U (12 of 14 invertebrates)]. In all six flowering plants, the helix is extended by one C-G base pair, thereby reducing the apex to four rather than six bases. The flagellates in general and *Giardia* in particular have rather different primary sequences, yet a 21-7 helix can still be constructed.

Despite the above, two other lines of evidence prevent the construction of a fully paired helix 21-7. First, although there is strong base pairing at both ends of the helix, there is a tendency for a central interior loop and this weak area occurs at different sites in different taxonomic groups. These include: site 3 [47 of 70 taxa with unpaired A, A]; site 4 [unpaired C, A (vascular plants and algae)]; site 5 [unpaired A opposite C (vascular plants and green algae)]; site 6 [unpaired A opposite A or C (amoeboid-like organisms and flagellates) or C (U) opposite U (vertebrates and invertebrates)]; and site 7 [C opposite A, A (13 of 14 vertebrates)]. Second, the 3′ side of the helix stem contains a group specific, additional base that must be placed at different sites to maximize base pairing. This base occurs between sites 5 and 6 (a U in chordates and most invertebrates), 2 and 3 (vascular plants), 3 and 4 (algae), or 4 and 5 (fungi and ciliates). In the other protists, the site of the extra base is variable. In the invertebrates, a one base slide in the 5′ direction repositions the U from site 8 to site 9 thereby resulting in a symmetrically paired helix.

It is the inability to use the same configuration across all taxonomic groups that has prompted DeWachter's group (2,15) to put forth the idea of a pseudoknot involving the 3′ strand of our proposed 21-7 in combination with the apex loop of helix 21-8. For the majority of taxa, a pseudoknot can be constructed with only occasional mismatches and/or a bulged base. There are apparent compensating changes, but an analysis of those changes does not strongly favor one model or the other. For example, *Giardia* has a very different sequence for the 3′ strand of helix 21-7, but compensating changes are found both on the 5′ strand of our helix 21-7 and in the apex loop of 21-8. In addition, a loss of helical structure is often evident, especially in the protists, at the four base pairs at the 3′ end of the pseudoknot . Furthermore, it is not clear how to accomodate the large 21-8 apical loop of *Plasmodium falciparum* C gene in the pseudoknot model. A combination of the two concepts would be to use our helix 21-6 and the pseudoknot thereby yielding an interstitial (unpaired) span of ca. 16 bases.

Helix 21-8

The stem of helix 21-8 is generally about 11 base pairs in length, contains few asymmetrical unpaired bases, and has a high frequency of compensating changes. For these reasons, it was discovered early by Olsen et al. (34). The apex loop has a strong signature of UUNUGUUGG and generally contains about 15 bases, although exceptions do occur, e.g. 24 to 34 bases in *Plasmodium falciparum*, *Euplotes*, and *Acanthamoeba*. The apex loop in *Giardia* is different from all known eukaryotes (CGCGCCGCGG), but these changes are compensated for in both the pseudoknot and 21-7 helix models. Most of this signature region is used as the 3′ strand of helix 21-7 in the pseudoknot (2). We have noted that the alignments for *Plasmodium berghei*, *Trypanosoma*, *Crithidia*, and *Euglena* (1) and for *Plasmodium falciparum* C gene, *Naegleria*, and *Leishmania* (2) are not in register for this particular sector based upon the assumption that the strong signature found in this apex loop is evidence of homology.

Helix 21-9

Four of the six flagellates have a large insertion between helices 21-8 and 22. Dams et al. (1) have folded these bases into one large helix (their helix 21-7; now 21-9) that contains several large asymmetrical internal loops. We propose that these bases can be incorporated into two shorter helices (21-9a and 21-9b) containing more base pairs as shown in Figure 4. These helices are still long and somewhat variable in length among the four taxa. There is clear evidence for homology in helix 21-9a among all four taxa, particularly between *Leishmania* and *Crithidia*. Although *Euglena* is more divergent than the former taxa regarding this helix, it shares a sequence within the apex loop (AUCGUU) with *Crithidia*. This sequence represents a more general signature (WUCGY) found in all four flagellates with this large insertion. *Naegleria* has a truncated 21-9a helix, but retains the sequence GAGCU at the 5′ base of the helix in common with *Crithidia* and *Leishmania*.

There is strong similarity between the 21-9b helices in *Crithidia* and *Leishmania* in both the stem elements and the apical loop. The equivalent helices of *Trypanosoma* and *Euglena* are very different from each other and from the previous two flagellates.

CONCLUDING REMARKS

Through the efforts of a large number of investigators, the elucidation of the secondary structure of the highly variable V4 region has progressed to the point that it is no longer necessary to depict this region as an undefined block of bases. Barring the discovery of an organism with a V4 region even more unusual than those already known, it should be possible to incorporate small subunit rRNA sequences of new organisms into existing alignments with a high degree of confidence. Both our analysis and the different, but complementing one of Neefs and DeWachter (15) suggest that helices 21-1b, 21-3, 21-5 and 21-8 deserve a high degree of confidence.

Areas common to all eukaryotes that will benefit from additional data include helix 21 and the pseudoknot area of helices 21-6 and 21-7. These areas show such high conservation that conclusive arguments relating to particular models can not be made at present. In contrast, the high variability in the large, insertion helices (21-1a, 21-2, 21-4 and 21-9) found primarily in the lower protists has precluded definitive conclusions regarding the number and structure of these helices. As additional sequences become available for the insects and protists, the 21-1/21-2 region should be understood with more confidence, as should the 21-9 region as more flagellate sequences are published. The use of compensating base changes to study these problems will hopefully be complemented by chemical analysis, direct visualization, directed mutational studies and other new approaches. In fact, it is conceivable that analysis by compensating base changes will be insufficient to prove secondary-structure models for the 'lead-in' (helix 21) and pseudoknot (helices 21-6/7) regions.

Our study of the compensating base changes in this V4 region has, in general, yielded phylogenetic groupings similar to the many recently published trees based on 18S rDNA sequences (e.g.6, 25, 35). We have been impressed with the degree of similarity between the green algae and the vascular plants (as might be expected) and between the green plants and the ascomycetes (as might not be expected). The relatedness of *Crithidia*, *Leishmannia* and *Trypanosoma* in comparison to the other flagellates is also conspicuous. As additional features of the V4 region become apparent, comparative studies that utilize higher order structure (helix number, helix length, base pairs within the helix, and constraints on helical vs. nonhelical regions) promise to further refine phylogenetic analyses.

ACKNOWLEDGEMENTS

The advice and valuable suggestions of Laurie Achenbach, Kirk Moberg, Gary Olsen and David Stock resulted in improvements in the final manuscript. Facilities provided by Carl Woese and instruction from his personnel, Robin Gutell and Will Weisburg in particular, were instrumental for the introduction of MLS to computer assisted analysis of sequences in the early phases of this project. This work was partially supported by NSF research grant (BSR 8918385 to DLN) and a NASA grant to Carl Woese. We acknowledge the support of the School of Life Sciences Center for Molecular Phylogeny (supported by the University of Illinois Research Board and NSF BBS 87-14603).

REFERENCES

1. Dams, E., Hendriks, L., Van de Peer, Y., Neefs, J.-M., Smits, G., Vandenbempt, I. and De Wachter, R. (1988) *Nucleic Acids Res.*, **16** (Suppl.), r87–r173.
2. Neefs, J.-M., Van de Peer, Y., Hendriks, L. and De Wachter, R. (1990) *Nucleic Acids Res.*, **18** (Suppl.), 2237–2317.
3. Woese, C.R., Magrum, L.J., Gupta, R., Siegel, R.B., Stahl, D.A., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J.J. and Noller, H.F. (1980) *Nucleic Acids Res.* **8**, 2275–2293.
4. Brimacombe, R. (1980) Biochem. Int., **1**,162–171.
5. Nelles, L., Fang, B.L., Volckaert, G., Vandenberghe, A. and De Wachter, R. (1984) *Nucleic Acids Res.*, **12**, 8749–8768.
6. Sogin, M.L., Gunderson, J.H., Elwood, H.J., Alonso, R.A. and Peattie, D.A. (1989) *Science*, **243**, 75–77.
7. Vossbrinck, C.R., Maddox, J.V., Friedman, S., Debrunner-Vossbrinck, B.A. and Woese, C.R. (1987) *Nature*, **326**, 411–414.
8. Woese, C.R., Gutell, R., Gupta, R. and Noller, H.F. (1983) *Microbiol. Rev.*, **47**, 621–69.
9. Gutell, R. R., Weiser, B., Woese, C.R. and Noller, H.F. (1985) *Progress Nucleic Acid Res. Mol. Biol.*, **32**, 155–216.
10. Zwieb, C., Glotz, C. and Brimacombe, R. (1981) *Nucleic Acids Res.*, **9**, 3621–3640.
11. Atmadja, J. and Brimacombe, R. (1984) *Nucleic Acids Res.*, **12**, 2649–2667.
12. Ellis, R.E., Sulston, J.E., and Coulson, A.R. (1986) *Nucleic Acids Res.*, 14, 2345–2364.
13. Gonzalez, I.L. and Schmickel, R.D. (1986) *Am. J. Hum. Genet.*, **38**,419–427.
14. Huysmans, E. and De Wachter, R. (1986) *Nucleic Acids Res.*, **14** (Suppl.), r73–r118.
15. Neefs, J.-M. and De Wachter, R. (1990) *Nucleic Acids Res.*, **18**, 5695–5704.
16. Mishler, B.D., Bremer, K, Humphries, C.J. and Churchill, S.P. (1988) *Taxon*, **37**, 391–395.
17. Steele, K.P., Holsinger, K.E., Jansen, R.K. and Taylor, D.W. (1988) *Taxon*, **37**, 135–138.
18. Olsen, G.J. (1988) Phylogenetic analysis using ribosomal RNA. *Methods Enzymol.*, **164**, 793–812.
19. Wheeler, W.C. and Honeycutt, R.L. (1988) *Mol. Biol. Evol.*, **5**, 90–96.
20. Hasselman, T., Camp, D.G. and Fox, G.E. (1989) *Nucleic Acids Res.*, **17**, 2215–2221.
21. Gutell, R.R. and Woese, C.R. (1990). *Proc. Natl. Acad. Sci. USA*, **87**, 663–667.
22. Waugh, D.S., Green, C. J. and Pace, N.R. (1989) *Science*, **244**, 1569–1571.
23. Hedges, B., Moberg, K.D., and Maxson, L.R. (1990). *Mol. Biol. Evol.*, **7**, 607–633.
24. Stock, D.W., Moberg, K.D., Maxon, L. R., and Whitt, G.S. (1990) *Environ. Biol. Fishes* (in press).
25. Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., and Raff, R.A. (1988) *Science*, **239**, 748–753.
26. Förster, H., Coffey, M.D., Elwood, H., and Sogin, M.L. (1990) *Mycologia*, **82**, 306–312.
27. Nickrent, D.L. and Franchina, C.R. (1990) *J. Mol. Evol.*, **31**, 294–301.
28. Bhattacharya, D. and Druehl, L.D. (1988) *J. Phycol.*, **24**, 539–543.
29. Herzog, M. and Maroteaux, L. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 8644–8648.
30. Waters, A.P. and McCutchan, T.F. (1989) *Nucleic Acids Res.*, **17**, 2135.
31. Zucker, M., and Stiegler, J. (1981) *Nuc. Acids. Res.*, **9**, 133–148.
32. Hendriks, L., DeBaere, R., Van Broeckhoven, C., and DeWachter, R. (1988) *FEBS Letters*, **232**, 115–120.
33. Tautz, D., Hancock, J.M., Webb, D.A., Tautz, C. and Dover, G.A. (1988) *Mol. Biol. Evol.*, **5**, 366–376.
34. Olsen, G.J., McCarroll, R., and Sogin, M. (1983) *Nucleic Acids Res.*, **11**, 8037–8049.
35. Cedergren, R., Gray, M.W., Abel, Y., and Sankoff, D. (1988) *J. Mol. Evol.*, **28**, 98–112.