The unusually long small subunit ribosomal RNA of *Phreatamoeba balamuthi*

Gregory Hinkle, Detlef D.Leipe, Thomas A.Nerad[1] and Mitchell L.Sogin*
Center for Molecular Evolution, Marine Biological Laboratory, Woods Hole, MA 02543 and
[1]Protistology Department, American Type Culture Collection, Rockville, MD 20852, USA

Received August 30, 1993; Revised and Accepted December 6, 1993 GenBank accession no. L23799

ABSTRACT

The small subunit ribosomal RNA (rRNA) of the anaerobic amoeba *Phreatamoeba balamuthi* is the longest 16S-like rRNA sequenced to date. Secondary structure analysis suggests that the additional sequence is incorporated in canonical eukaryotic expansion regions and is not due to the presence of introns. Reverse transcriptase sequencing of total RNA extracts confirmed that two uncommonly long expansion regions are present in native *P.balamuthi* 16S-like rRNA. Primary sequence comparison and similar secondary structure indicate a 61 base stem and loop repeat within an expansion region; a mechanism whereby the repeat may have been incorporated is presented. *P.balamuthi* provides further evidence that 16S-like rRNA length does not correlate with phylogenetic position.

INTRODUCTION

Widely used as molecular chronometers, 16S-like rRNA has rapidly become one of the largest and phylogenetically most diverse molecular databases [1]. Eukaryotic 16S-like rRNAs are typically 1800 bases in length, but molecules as short as 1246 bases (the microsporidian *Vairimorpha necatrix* [2]) and as long as 2469 bases (the pea aphid, *Acyrthosiphon pisum* [3]) have been described. Small subunit rRNA can be considered a mosaic of genetic elements with varying degrees of sequence divergence; well conserved regions are interspersed with moderately to hypervariable regions. All known 16S-like rRNAs can be folded into consensus secondary structures containing approximately 50 helices [4]. Length variation among 16S-like rRNA molecules is generally limited to expansion regions outside of highly conserved 'core' sequences. Length variation within conserved domains is limited to mobile introns and as such is removed during ribosome assembly [5]. Universally conserved regions within 16S-like rRNAs is evidence that essential secondary and tertiary structures were established in the common ancestor to archaebacteria, eubacteria and eukaryotes and have been maintained for eons [6,7]. Sogin and Edman [5] have argued that in most cases the only apparent constraint on secondary structure in variable regions may be the maintenance of ribosomal function, i.e., stems and loops must not disrupt translation. Gunderson and Sogin [8] demonstrated the stage/host specific transcription of structurally distinct 16S-like rRNA genes in *Plasmodium*. Differences between the *Plasmodium* sequences are primarily restricted to hypervariable regions defining long helices. Extremes in length variation are of interest then in defining the constraints of ribosomal structure and function.

MATERIALS AND METHODS

A clonal subculture of *Phreatamoeba balamuthi* (ATCC # 30984) was grown in liquid ATCC medium 114 at 35°C and harvested by centrifugation at 4°C. Total nucleic acids were isolated and the 16S-like nuclear coding regions were selectively amplified using polymerase chain reaction (PCR) techniques as described by Medlin [9]. Amplified rRNA was cloned into M13 sequencing phage. Multiple clones were pooled and both strands sequenced in entirety with Sequenase 2.0 (US Biochemical) using synthetic oligoprimers complementary to phylogenetically conserved 16S-like rRNA sequences [10].

Highly conserved regions within the 16S-like rRNA sequences were aligned with all known eukaryotic as well as representative eubacterial and archaebacterial 16S-like sequences [1] through the introduction of gaps. The alignment was refined by consideration of phylogenetically conserved higher order structures; regions defining proven helices were juxtaposed. Only those positions that were in obvious alignment (using the criteria of primary and/or secondary structure conservation) were used in the phylogenetic analyses (1023 positions; data available upon request from the authors). The organisms in the analysis were *Acanthamoeba castellanii* (GenBank M13435), *Achlya bisexualis* (GenBank M32705), *Bacteroides fragilis* (GenBank M11656), *Chlamydomonas reinhardtii* (GenBank M32703), *Costaria costata* (GenBank M97958), *Crypthecodinium cohnii* (GenBank M64245), *Diaphanoeca grandis* (GenBank L10824), *Dictyostelium discoideum* (GenBank K02641) *Encephalitozoon cuniculi* (GenBank Z19563), *Escherichia coli* (GenBank J01859), *Emiliania huxleyi* (GenBank L04957), *Entamoeba histolytica* (GenBank X61116) *Euglena gracilis* (GenBank M12677), *Gracilaria lemaneiformis* (GenBank M54986), *Hexamita inflata* (GenBank L07836), *Microciona prolifera* (GenBank L10825),

*To whom correspondence should be addressed

Mnemiopsis leidyi (GenBank L10826), *Naegleria gruberi* (GenBank M18732), *Ochromonas danica* (GenBank M32704), *Oxytricha nova* (GenBank M14601), *Physarum polycephalum* (GenBank X13160), *Prorocentrum micans* (GenBank M14649), *Saccharomyces cerevisiae* (GenBank J01353), *Schizosaccharomyces pombe* (GenBank X54866), *Sulfolobus solfataricus* (GenBank X03235), *Tritrichomonas foetus* (GenBank M81842), *Thermoplasma acidophilum* (GenBank M38637), *Trypanosoma brucei* (GenBank M12676), and *Zea mays* (GenBank K02202). (GenBank K02202). Sequence similarities were converted to

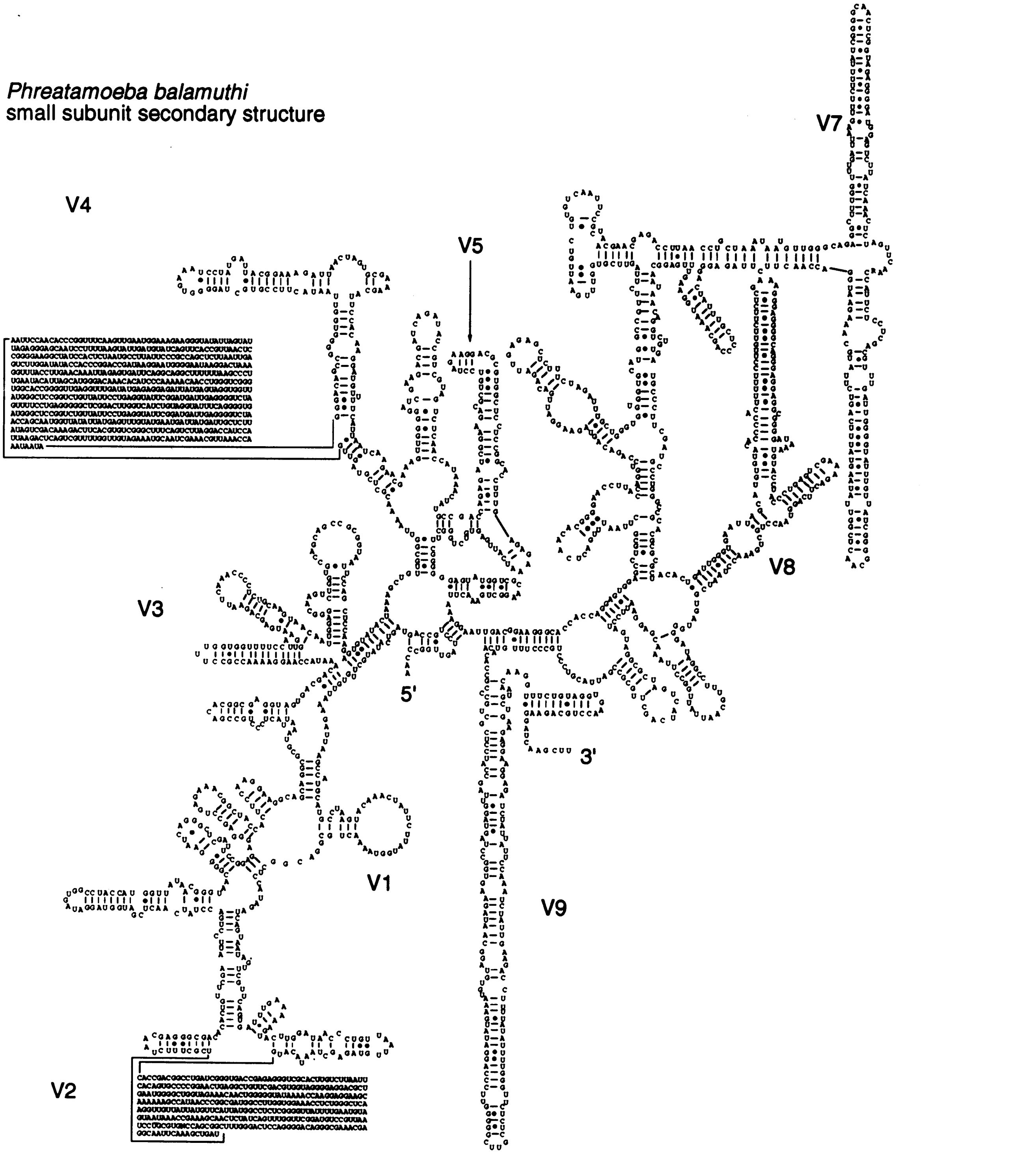

Figure 1. Secondary structure model of the 16S-like rRNA of *Phreatamoeba balamuthi*. The sequence was fitted to previously published secondary structure models [15]. The regions labelled V1–9 indicate variable regions as defined by Neefs, *et al.* [20]. Canonical base pairs are indicated with a dash (–), noncanonical base pairs (U-G) with a filled circle ●.

structural distances expressed in nucleotide changes/site by the method of Jukes and Cantor [11]. Phylogenetic trees were inferred by the method of Olsen [12]. Maximum likelihood analyses were performed using the program fastDNAml [13].

Two *P.balamuthi* expansion zones were sequenced in their entirety using the bulk extracted nucleic acids by the method of Carpenter and Simon [14]. Secondary structure for *P.balamuthi* 16S-like rRNA was drawn using previously published eukaryotic structures [15] as models.

RESULTS

Primary and secondary structure

The complete 16S-like rRNA sequence of *P.balamuthi* has been submitted to GenBank (accession # L23799). The 16S-like rRNA molecule is 2741 bases long and has a 47% G+C content. Direct inspection as well as BLAST [16] searches of relevant databases did not find evidence for intron signature sequences [17]. To verify that the nuclear genomic sequence is present in native ribosomes, we sequenced two unusually large *P.balamuthi* expansion regions, positions 264–405 and 1441–1580. Despite several ambiguous positions, reverse transcriptase sequencing demonstrated the presence of the two expansion regions (data not shown). A consensus secondary structure of the molecule is in Figure 1. The variable regions (V1–V9) are numbered following the model of Neefs *et al.* [18]. All the universal primary sequences and core helices are present in *P.balamuthi* 16S-like rDNA. Though *P.balamuthi* expansion stem and loop structures are not proven, the potential for extensive base pairing strongly resembles similar structures proposed for other eukaryotic 16S-like rRNA [19].

Phylogenetic analysis

An unrooted, calculated distance tree of *P.balamuthi* and representative eukaryotes is presented in Figure 2. Maximum likelihood analysis of the same dataset gave rise to a nearly identical tree topology (data not shown).

DISCUSSION

The exceptional length of *P.balamuthi* 16S-like rRNA is due to atypically large expansion zones. Particularly within the hypervariable zones V2 and V4, as well as in the moderately variable V7, V8 and V9, the additional sequence is unusually long (Figure 1). In all other respects *P.balamuthi* 16S-like RNA is typical; all of the canonical eukaryotic core elements are present and there is no evidence of a novel, large expansion zone. Unlike *Acanthamoeba castellanii*, another amoeba with an unusually long 16S-like rRNA (2303 bases), there is no evidence for expansion in region V5 [20]. Although the entire molecule was not sequenced with reverse transcriptase, the confirmation that two unusually long expansion regions (264–405 and 1441–1580) are present in native ribosomes is strong evidence that the entire genomic transcript remains intact.

Helices are considered conserved if they are found in the majority of 16S-like rRNAs. Several compensating base changes that maintain secondary structure are taken as evidence that a helix exists *in vivo* [6]. That most 16S-like rRNA expansion regions can be folded in a series of helices suggests that expansion regions without secondary structure may disrupt pre-existing secondary structures and as such are strongly selected against [19]. Whether or not expansion regions have a functional role in ribosome assembly or in translation is not known. Gerbi [21]

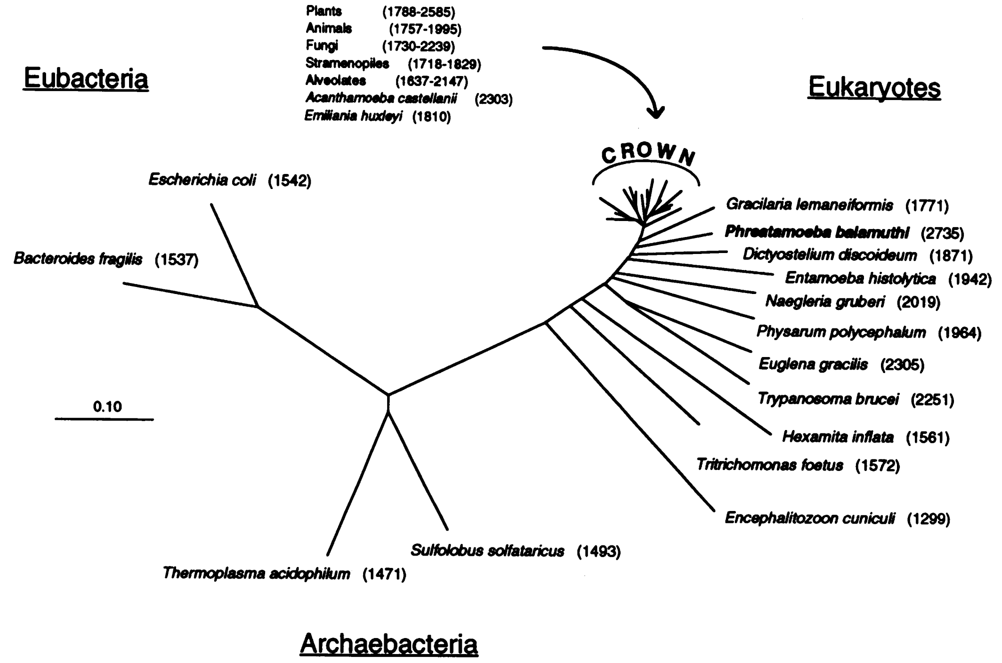
Eubacteria
Plants (1788-2585)
Animals (1757-1995)
Fungi (1730-2239)
Stramenopiles (1718-1829)
Alveolates (1637-2147)
Acanthamoeba castellanii (2303)
Emiliania huxleyi (1810)
Eukaryotes
CROWN
Escherichia coli (1542)
Bacteroides fragilis (1537)
Gracilaria lemaneiformis (1771)
Phreatamoeba balamuthi (2735)
Dictyostelium discoideum (1871)
Entamoeba histolytica (1942)
Naegleria gruberi (2019)
Physarum polycephalum (1964)
Euglena gracilis (2305)
Trypanosoma brucei (2251)
Hexamita inflata (1561)
Tritrichomonas foetus (1572)
Encephalitozoon cuniculi (1299)
0.10
Sulfolobus solfataricus (1493)
Thermoplasma acidophilum (1471)
Archaebacteria

Figure 2. An unrooted universal phylogeny derived from a distance matrix analysis of aligned, representative, eukaryotic 16S-like rRNA sequences. The scale bar corresponds to 10 changes per 100 nucleotide positions. The length in nucleotides of the complete 16S-like rRNA paranthetically follows each organism.

Table 1. Length in nucleotides of variable regions V1–V9 of the representative eukaryotes in Figure 2

ORGANISM	V1	V2	V3	V4	V5	V7	V8	V9
Phreatamoeba balamuthi	29	452	61	657	52	175	129	88
Mnemiopsis leidyi	28	196	58	225	41	28	59	70
Microciona prolifera	30	202	58	223	41	27	57	75
Diaphanoeca grandis	30	184	58	229	38	27	61	72
Caenorhabditis elegans	19	206	68	217	48	24	58	114
Schizosaccharomyces pombe	30	195	59	236	43	30	61	73
Saccharomyces cerevisiae	29	193	59	224	40	27	78	72
Blastocladiella emersonii	31	222	69	260	63	28	103	114
Zea mays	30	196	59	225	42	29	58	74
Chlamydomonas reinhardtii	29	192	58	222	42	27	56	68
Acanthamoeba castellanii	30	270	61	370	132	131	117	92
Dictyostelium discoideum	28	190	56	226	41	110	56	67
Gracilaria lemaneiformis	29	171	58	221	41	33	53	70
Oxytricha nova	27	192	55	219	38	26	57	59
Prorocentrum micans	29	196	57	222	42	29	60	69
Crypthecodinium cohnii	28	192	57	224	41	30	59	68
Plasmodium berghei	29	202	69	276	89	156	111	111
Emiliania huxleyi	28	191	58	225	42	27	59	72
Achlya bisexualis	30	189	58	232	43	29	59	72
Ochromonas danica	29	183	57	226	40	31	57	70
Costaria costata	29	195	58	236	43	30	63	70
Entamoeba histolytica	29	189	58	210	87	52	116	53
Naegleria gruberi	44	252	74	308	43	83	71	52
Physarum polycephalum	31	258	55	293	38	54	67	76
Euglena gracilaria	28	293	62	516	52	100	75	77
Trypanosoma brucei	26	262	63	398	168	88	83	58
Tritrichomonas foetus	20	133	52	114	31	25	55	44
Hexamita inflata	28	138	49	110	28	0	57	56
Encephalitozoon cuniculi	23	81	25	62	23	4	21	27

suggested that expansion regions have no essential function and are tolerated so long as they are not disruptive. Comparative 16S-like rRNA has identified structural constraints involving tetra loops, hairpin loops of 4 nucleotides, most typically GNRA, UNCG or CUUG, that are known to add additional stability [22]. The presence of tetra loop sequences capping several *P.balamuthi* helices (Figure 1) may prevent the establishment of secondary structure between hypervariable and conserved regions. As the longest 16S-like rRNA sequenced to date, *P.balamuthi* establishes a new upper boundary for the number of bases that fold in such a way that ribosomal function is maintained.

In the process of constructing the secondary structure, we recognized the presence of a 61 base repeat within the V7 expansion region (Figure 3). After the introduction of 3 gaps, there is 80% similarity between the regions 2087–2146 and 2171–2232, including a stretch of 18 identical nucleotides. As can be seen in Figure 1, these two regions form adjacent stem and loops structures within the V7 expansion zone. Incorporation of length variation in expansion zones is not well understood. Whether expansion occurs cumulatively from single events or collectively in relatively large scale insertions is relevant to reconstruction of evolutionary history through comparison of molecular sequences. Short (1–5n), repetitive nucleotide repeats are common in expansion zones and are prominent in the V7 region of the heretofore longest 16S-like rRNA molecule of *Acyrthosiphon pisum* [3]. Simple, repetitive sequence motifs are most likely the result of slippage events during replication [23]. There is no evidence of short nucleotide motifs within the repeat stem and loop structures of *P.balamuthi*; the scale of the repeat (61 bases) suggests that slippage is not limited to short motifs or that length variation in hypervariable regions can arise by multiple, large insertion events. A large scale slippage event is modeled in Figure 4. If a stable stem and loop structure were to form on only one strand during DNA replication, a repeat could in principle become incorporated in a genomic copy of a 16S-like rDNA gene. After replication, the stability of the stem and loop structure would maintain the noncanonical form (where one strand has a single copy and the other strand two copies of the repeat). During subsequent DNA replication, the repeat would become incorporated in the complementary strand. Selection against the introduction of additional secondary structure within the ribosome would likely be directly related to the degree that translation would be disrupted [24]. As a stable stem and loop structure, the repeat in *P.balamuthi* may have been predisposed for incorporation in an expansion zone.

```
2082 5'-GUUUGAUU-AAGUUUCUUUUAUCGGGGCAACUCGGU-AGAGGGAUGGAGUCUUAU-CA-3' 2138
2166 5'-GUUUGAUUGGGGUAUUUGUUAUCGGGGCAACUCGGUUAUAAGUAUUUCAGUCUUAUGCA-3' 2224
```

Figure 3. Aligned sequences making up a repeat in the V7 region of the 16S-like rRNA of *P.balamuthi*. The absolute position numbers from the 5′ end of the 16S-like rRNA molecule flank the sequences. Three alignment gaps have been introduced in the upper sequence. Identical bases are in bold. A stretch of 18 identical bases with no length variation is underlined.

P.balamuthi adds additional support that neither overall 16S-like rRNA length (Figure 2) nor individual variable (Table 1) regions correlate with phylogenetic position [15]. Although the early branching microsporidia have unusually short 16S-like rRNA, shorter than either eubacteria or archaebacteria, there is no evidence for an increase in length along the eukaryotic line of descent. As can be seen in Figure 2, unusually long 16S-like rRNAs can be found among many evolutionarily distant lineages.

Molecular phylogenetic analyses suggest that the ancestor of the eukaryotic lineage was an anaerobe [25–27]. Because of their comparative morphological simplicity (no mitochondria, no golgi apparatus) anaerobic amoeba such as *P.balamuthi* are expected to represent an early diverging lineage in the eukaryotic line of

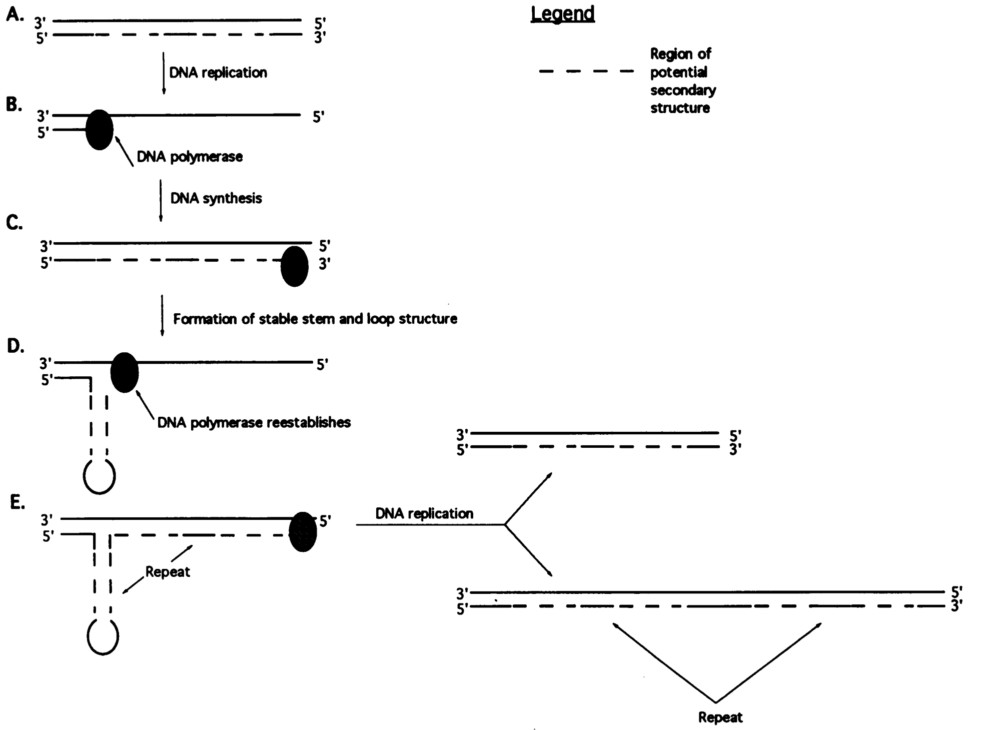

Figure 4. A model for the introduction and establishment of the repeat sequence identified in region V7 of the 16S-like rRNA of *P.balamuthi*. **A.** Strand separation during DNA replication. **B.** DNA polymerase binding. **C.** Synthesis of complementary strand. **D.** Establishment of stem and loop structure on one strand and reestablishment of DNA polymerase E. Second synthesis of complementary strand and segregation during the subsequent replication.

descent. The phylogeny depicted in Figure 2 demonstrates that *P.balamuthi* diverges before the 'crown' [28], but relatively late among eukaryotes; mitochondria and the golgi apparatus were very likely secondarily lost. Although they all form a cyst and have a flagellated stage, vahlkampfiid amoebae (e.g., *N.gruberi*) and *P.balamuthi* are clearly not close relatives. Furthermore our data reaffirm that amoeba are highly polyphyletic [24, 29]. *P.balamuthi* does not group with any previously sequenced amoeba, or other eukaryote. Until more amoebae 16S-like rRNA are sequenced the genetic diversity of the amoeba phenotype will remain obscure.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (Le 743/1-1 to DLL), the G.Unger Vetleson Foundation and the National Institutes of Health (to MLS).

REFERENCES

1. Larsen,N., Olsen,G.J., Maidak,B.L., McCaughey,M.J., Overbeek,R., Macke,T.J., Marsh,T.L and Woese,C.R. (1993) *Nucl. Acids Res.* **21**, 3021–3023.
2. Vossbrinck,C.R., Maddox,J.V., Friedman,S., Debrunner-Vossbrinck,B.A. and Woese,C.R. (1987) *Nature* **326**, 411–414.
3. Kwon,O., Ogino,K. and Ishikawa, H. (1991) *Eur. J. Biochem.* **202**, 827–833.
4. Guttell, R.R., Weiser,B., Woese,C.R. and Noller,H.F. (1985) *Prog. Nucl. Acids. Res. Mol. Biol.* **32**, 155–216.
5. Sogin,M.L. and Edman,J.C. (1989) *Nucl. Acids Res.* **17**, 5349–5359.
6. Sogin,M.L. and Gunderson,J.H. (1987) *Ann. N. Y. Acad. Sci.* **503**, 125–39.
7. Gray,M.W., Sankoff,D. and Cedergren,R.J. (1984) *Nucl. Acids Res.* **12**, 5837–5852.
8. Gunderson,J.H. and Sogin,M.L., Wollett,G., Holingdale, M., de la Cruz, V.V., Waters, A.P. and McCutchan, T.F. (1987) *Science* **238**, 933–937.
9. Medlin,L., Elwood,H.J., Stickel,S. and Sogin,M.L. (1988) *Gene* **71**, 491–499.
10. Elwood,H.J., Olsen,G.J. and Sogin,M.L. (1985) *Mol. Biol. Evol.* **2**, 399–410.
11. Jukes,T.H. and Cantor,C.R. (1969) In Munro,H.N. (ed.), Mammalian Protein Metabolism. Academic Press, New York, New York, pp. 21–32.
12. Olsen,G.J. (1988) *Methods Enzymol.* **164**, 793–812.
13. Olsen,G.J., Matsuda,H., Hagstrom,R. and Overbeek,R. (1993) *Cabios*. In press.
14. Carpernter,C.D. and Simon,A.E. (1990) *BioTechniques* **8**, 26–27.
15. Gutell,R.R. (1993) *Nucl. Acids Res.* **21**, 3051–3054.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.* **215**, 403–10.
17. Cech,T.R. (1988) *Gene* **73**, 259–71.
18. Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucl. Acids Res.* **21**, 3025–3049.
19. Hancock,J.M. and Dover,G.A. (1988) *Mol. Biol. Evol.* **5**, 377–391.
20. Gunderson,J.G. and Sogin,M.L. (1986) *Gene* **44**, 63–70.
21. Gerbi,S.A. (1986) BioSystems **19**, 247–258.
22. Woese,C.R. Winker,S. and Gutell,R.R. (1990) *Proc. Natl. Acad. Sci.* **87**, 8467–8471.
23. Dover,G.A. and Hancock,J.M. (1990) *Nucl. Acids Res.* **18**, 5949–5954.
24. Sogin,M.L. (1991) *Curr. Opin. Genet. Dev.* **1**, 457–463.
25. Leipe,D.D., Gunderson,J.,G., Nerad, T.A. and Sogin,M.L. (1993) *Mol. Biochem. Parisitol.* **59**, 41–49.
26. Patterson,D.J. and Sogin,M.L. (1993) In Hartman,H. and Matsuno,K. (eds.), The Origin and Evolution of Prokaryotic and Eukaryotic Cells. World Scientific, New Jersey, pp. 13–46.
27. Van Keulen,H., Gutell,R.R., Gates,M.A., Campbell,S.R., Erlandsen,S.L., Jarroll,E.L., Kulda,J. and Meyer,E.A. (1993) *FASEB* **7**, 223–231.
28. Knoll,A.H. (1992) *Science* **256**, 622.
29. Hinkle,G. and Sogin,M.L. (1993) *J. Euk. Micro.* **1**, 599–603.