doi:10.1006/jmbi.2000.3529 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2000) **297**, 365–380

JMB

Complete Sequence of the Mitochondrial Genome of *Tetrahymena pyriformis* and Comparison with *Paramecium aurelia* Mitochondrial DNA

Gertraud Burger[1], Yun Zhu[1], Tim G. Littlejohn[1], Spencer J. Greenwood[2] Murray N. Schnare[2], B. Franz Lang[1] and Michael W. Gray[2]*

[1]*Program in Evolutionary Biology, Canadian Institute for Advanced Research Département de Biochimie Université de Montréal Montréal, Québec, Canada H3C 3J7*

[2]*Program in Evolutionary Biology, Canadian Institute for Advanced Research Department of Biochemistry and Molecular Biology Dalhousie University, Halifax Nova Scotia, Canada, B3H 4H7*

**Corresponding author*

We report the complete nucleotide sequence of the *Tetrahymena pyriformis* mitochondrial genome and a comparison of its gene content and organization with that of *Paramecium aurelia* mtDNA. *T. pyriformis* mtDNA is a linear molecule of 47,172 bp (78.7 % A + T) excluding telomeric sequences (identical tandem repeats of 31 bp at each end of the genome). In addition to genes encoding the previously described bipartite small and large subunit rRNAs, the *T. pyriformis* mitochondrial genome contains 21 protein-coding genes that are clearly homologous to genes of defined function in other mtDNAs, including one (*yejR*) that specifies a component of a cytochrome *c* biogenesis pathway. As well, *T. pyriformis* mtDNA contains 22 open reading frames of unknown function larger than 60 codons, potentially specifying proteins ranging in size from 74 to 1386 amino acid residues. A total of 13 of these open reading frames (''ciliate-specific'') are found in *P. aurelia* mtDNA, whereas the remaining nine appear to be unique to *T. pyriformis*; however, of the latter, five are positionally equivalent and of similar size in the two ciliate mitochondrial genomes, suggesting they may also be homologous, even though this is not evident from sequence comparisons. Only eight tRNA genes encoding seven distinct tRNAs are found in *T. pyriformis* mtDNA, formally confirming a long-standing proposal that most *T. pyriformis* mitochondrial tRNAs are nucleus-encoded species imported from the cytosol. Atypical features of mitochondrial gene organization and expression in *T. pyriformis* mtDNA include split and rearranged large subunit rRNA genes, as well as a split *nad1* gene (encoding subunit 1 of NADH dehydrogenase of respiratory complex I) whose two segments are located on and transcribed from opposite strands, as is also the case in *P. aurelia*. Gene content and arrangement are very similar in *T. pyriformis* and *P. aurelia* mtDNAs, the two differing by a limited number of duplication, inversion and rearrangement events. Phylogenetic analyses using concatenated sequences of several mtDNA-encoded proteins provide high bootstrap support for the monophyly of alveolates (ciliates, dinoflagellates and apicomplexans) and slime molds.

Keywords: ciliates; alveolates; protists; phylogeny; genetic map

Introduction

Alveolata is a recently recognized assemblage of unicellular eukaryotes (protists) that are characterized by the presence of cortical alveoli (cavities or pits in the outer envelope) and of mitochondria with tubular cristae. The alveolates comprise three major phyla; Ciliophora (ciliate protozoa), Dinozoa (dinoflagellates) and Apicomplexa (a mostly parasitic group of organisms that includes Plasmodium,

Present address: Tim G. Littlejohn, The Australian National Genomic Information Service (ANGIS), University of Sydney, New South Wales 2006, Australia.

Abbreviations used: mtDNA, mitochondrial DNA; ORF, open reading frame; LSU, large subunit; SSU, small subunit; RT, reverse transcriptase.

E-mail address of the corresponding author: M.W.Gray@Dal.Ca

0022-2836/00/020365–16 $35.00/0

the causative agent of malaria) (Patterson & Sogin, 1992; Cavalier-Smith, 1993). The evolutionary cohesion of this group is strongly supported in phylogenetic trees based on small subunit (SSU) rRNA sequence comparisons, with Dinozoa and Apicomplexa appearing as sister clades and Ciliophora branching more deeply (Sogin, 1989; Schlegel, 1991; Cavalier-Smith, 1993).

Despite the fact that the alveolates constitute a major group of protists encompassing a high degree of biological and phylogenetic diversity, we know relatively little about the structure, function and evolution of mitochondrial DNA (mtDNA) in these organisms. A 6 to 7 kb linear mitochondrial genome has been identified in four apicomplexan species (*Plasmodium yoelii*, *Plasmodium falciparum*, *Plasmodium vivax* and *Theileria parva*) and completely sequenced (Vaidya *et al*., 1989, 1993; Feagin *et al*., 1992; Kairo *et al*., 1994; McIntosh *et al*., 1998; Sharma *et al*., 1998). Although this DNA is thought to be the functional equivalent of the mitochondrial genome in other organisms, it is highly unusual in gene content and organization, being by far the smallest mtDNA yet described, with only three protein-coding genes, no tRNA genes, and fragmented and scrambled large subunit (LSU) and small subunit rRNA genes (Feagin, 1994; Wilson & Williamson, 1997). Two ciliate mtDNAs, from *Paramecium aurelia* and *Tetrahymena pyriformis* (classes Nassophorea and Oligohymenophorea, respectively), have been extensively characterized (work summarized by Cummings, 1992; Gray, 1992; Gray *et al*., 1998). These DNA species, while also linear, are some 40-50 kb in size, i.e. almost tenfold larger than their apicomplexan counterparts, and contain many more of the standard set of genes encoded by mtDNA in other organisms. At present, virtually nothing is known about mtDNA in Dinozoa, although the *cox1* gene of a non-photosynthetic dinoflagellate, *Crypthecodinium cohnii*, has recently been sequenced (Norman & Gray, 1997; Inagaki *et al*., 1997). Hence, we have at present only a fragmentary picture of mtDNA structure and function in the alveolates.

Ciliates have mitochondrial genomes whose sizes fall within the range (20-60 kb) more typical of protist mtDNAs (see Gray *et al*., 1998). Determination of the complete sequence of *P. aurelia* mtDNA (40,469 bp; Pritchard *et al*., 1990b) revealed a number of novel features, including the apparent absence of certain genes encoded by mtDNA in almost all other eukaryotes and the presence of an unusually high number of unassigned open reading frames (ORFs) (Cummings, 1992). Sequence comparisons indicated an exceptionally high rate of primary structure divergence in identified protein-coding genes relative to their homologs in other eukaryotes (Pritchard *et al*., 1990a), complicating an assessment of whether these unassigned ORFs are real genes and, if so, what their functions may be.

In order to better understand ciliate mtDNA diversity and evolution, we have determined the complete sequence of the mtDNA from a second ciliate protozoon, *T. pyriformis*. Although at the outset of this project some sequence information was available for mtDNA in Tetrahymena species (primarily *T. pyriformis*), these data were rather limited, representing mainly rRNA and tRNA genes and immediately flanking regions (Schnare *et al*., 1986; Heinonen *et al*., 1987, 1990; Suyama, 1985; Suyama *et al*., 1987; Labriola *et al*., 1987; Hekele & Beier, 1991) rather than protein-coding genes (Ziaie & Suyama, 1987; Suyama & Jenney, 1989). As we report here, availability of the complete *T. pyriformis* mtDNA sequence has permitted a more comprehensive and incisive analysis of gene content and organization in *P. aurelia* mtDNA than was previously possible with the latter sequence alone, and has provided information about how these two genomes have changed since their separation from a common ancestor.

Results

Physical properties and sequence of *T. pyriformis* mtDNA

Early studies indicated that the mtDNA of *T. pyriformis* strain ST is a linear duplex of length 17.6 μm having an estimated molecular mass of 33.8×10^6 Da (Suyama & Miura, 1968), in good agreement with other measurements suggesting a size of 28.4×10^6 Da (sedimentation analysis; Goldbach *et al*., 1977) to 30×10^6 Da (kinetic complexity; Flavell & Jones, 1970). For a DNA molecule having an A + T content of 75 % (Flavell & Jones, 1970; Goldbach *et al*., 1977), these values correspond to a size of about 45-50 kb. The complete sequence of the mtDNA of *T. pyriformis* strain ST reveals a basic genome size (excluding telomeres) of 47,172 bp and an A + T content of 78.7 %, substantially confirming these earlier conclusions.

Renaturation and rRNA-DNA hybridization studies have demonstrated that the mtDNA of *T. pyriformis* strain ST contains a near-terminal duplication-inversion, each duplicated region containing a large subunit rRNA gene (Arnberg *et al*., 1975; Goldbach *et al*., 1977, 1978a,b). In addition, terminal length heterogeneity has been observed and attributed to the presence of variable numbers of repeated elements (telomeres) at the ends of the linear mtDNA (Goldbach *et al*., 1977). These features have been characterized at the sequence level (Heinonen *et al*., 1987, 1990; Middleton & Jones, 1987; Morin & Cech, 1988a,b) and are confirmed here. The telomeric regions comprise direct tandem arrays of 31 bp repeats that are identical in sequence at the two ends of the chromosome, in contrast to what is seen in some other Tetrahymena species whose mtDNA has different telomeric repeats at the two ends (Morin & Cech, 1988a).

Sequences corresponding to the central *rns* gene and duplicated *rnl* gene have been determined (Schnare *et al*., 1986; Heinonen *et al*., 1987, 1990)

and were re-sequenced in their entirety in the course of the present study. Sequence differences at four positions, all within *rnl* coding regions, were found within 8826 bp of independently determined sequence. The significance of these discrepancies is discussed below. Other differences were noted in comparison with partial sequences for *T. pyriformis* strain ST mtDNA that have been published by Suyama and co-workers (Suyama, 1985; Suyama *et al*., 1987; Labriola *et al*., 1987; Ziaie & Suyama, 1987; Suyama & Jenney, 1989). At each site of disagreement, we re-checked and verified our assignment.

Outside of the *rnl* region, two polymorphic sites were found during random sequencing. Both of these sites are in protein-coding genes and both are silent (third-position) codon changes. One polymorphism (A or C at position 17,805 in GenBank acc. no. AF160864) is in *rpl16*, resulting in either ATA or ATC Ile codons; the other (C or T at position 44,059) is in an unidentified gene, *ymf73* (see below), resulting in alternate codons GCC or GCT (both Ala).

Gene content and overall organization

The *T. pyriformis* mtDNA contains genes for two rRNAs (large subunit and small subunit, each consisting of two separate pieces), only seven distinct tRNAs and 21 proteins of known function, as well as 22 unassigned ORFs of more than 60 codons whose functions remain to be defined (Table 1). The content of identified protein-coding genes is almost identical in *T. pyriformis* and *P. aurelia* mtDNAs (Table 2). Both genomes apparently lack several well-conserved respiratory chain genes (*nad4L*, *nad6*, *cox3*, *atp6*) that are encoded by animal and most other mtDNAs, whereas both contain additional NAD dehydrogenase subunit genes (*nad7*, *nad9* and *nad10*) as well as small subunit and large subunit ribosomal protein genes. The latter are a subset of the ribosomal protein genes found in some plant and protist mtDNAs (Table 2; see also Gray *et al*., 1998), with *rps19* present in *T. pyriformis* but not *P. aurelia* mtDNA. Both mtDNAs contain *yejR* (*ccl1*), a homolog of a gene in plant and some other protist mtDNAs that encodes a component of a cytochrome *c* heme lyase, probably the catalytic subunit (Thöny-Meyer, 1987). A novel organizational feature in both mtDNAs is the presence of a split *nad1* gene, whose two parts are encoded on different strands of the genome; the expression of this split gene is described in the accompanying paper (Edqvist *et al*., 2000). No intron sequence was previously reported in *P. aurelia* mtDNA by Pritchard *et al*. (1990b) and none was encountered in any of our analyses of the *T. pyriformis* or *P. aurelia* mtDNA sequences. Neither *rrn5* (which encodes 5 S rRNA) nor *rnpB* (which specifies the RNA subunit of RNase P) could be identified in either *T. pyriformis* or *P. aurelia* mtDNA.

Table 1. Genes identified in *Tetrahymena pyriformis* mtDNA

A. *Ribosomal RNA* (2)[a]
   Large subunit (*rnl_a*, *rnl_b*)[b], small subunit (*rns_a*, *rns_b*)
B. *Transfer RNA* (7)[d]
   trnE(uuc), *trnF(gaa)*, *trnH(gug)*, *trnL(uaa)*[c], *trnM(cau)*, *trnW(uca)*, *trnY(gua)*
C. *Electron transport/oxidative phosphorylation* (12)
   Respiratory chain (11)
      NADH dehydrogenase (*nad1*[e], *nad2*, *nad3*, *nad4*, *nad5*, *nad7*, *nad9*, *nad10*)
      Apocytochrome *b* (*cob*)
      Cytochrome oxidase (*cox1*, *cox2*)
   ATP synthase complex (1)
      F_0-ATPase (*atp9*)
D. *Ribosomal protein* (8)
   Small subunit (5): *rps3*, *rps12*, *rps13*, *rps14*, *rps19*
   Large subunit (3): *rpl2*, *rpl14*, *rpl16*
E. *Protein transport and maturation* (1)
      yejR (*ccl1*)
F. *Ciliate-specific ORFs*[f] (13)
      (see Table 3)
G. *ORFs unique to* T. pyriformis *mtDNA* (9)
      (see Table 4)

[a] Separate cistrons (*a* and *b*) encode the two portions of split LSU and SSU rRNAs (see Figure 1 and Schnare *et al*., 1986; Heinonen *et al*., 1987).

[b] Duplicate non-identical cistrons (see Figure 1 and Heinonen *et al*., 1990).

[c] Duplicate identical genes (see Figure 1 and Heinonen *et al*., 1987).

[d] Anticodon sequences are shown in lowercase letters in parentheses.

[e] N-terminal and C-terminal portions of the Nad1 protein are encoded by separate cistrons (*nad1_a* and *nad1_b*) located on opposite strands (see Figure 1).

[f] Unassigned homologous ORFs present in both *T. pyriformis* and *P. aurelia* mtDNAs, but not identified to date in the mitochondrial genome of any other organism.

On the basis of sequence comparisons (Table 3), 13 of the 22 unassigned ORFs in Tetrahymena mtDNA have evident homologs in *P. aurelia* mtDNA; however, these genes are not present in any other sequenced mitochondrial genome, as far as we can determine. We designate these 13 ORFs ''ciliate-specific''. In three instances (*ymf65*, *ymf66*, *ymf67*), two adjacent ORFs in Paramecium mtDNA appear to be equivalent to a single ORF in the Tetrahymena mitochondrial genome. Considering that Paramecium ORF13 (*orf90*, Table 3) would specify a protein of only 90 amino acid residues, compared with the 443 amino acid residues encoded by *ymf67* in *T. pyriformis* mtDNA, we suggest that ORF12 (*orf265_1*, Table 3), which is located immediately upstream of ORF13 (Figure 1), constitutes an additional portion (*ymf67_a*) of the *ymf67* equivalent in *P. aurelia* mtDNA, even though there is no evident similarity between the corresponding Tetrahymena and Paramecium sequences in this instance. In support of this conclusion, we note that re-sequencing has provided evidence that Paramecium ORFs 12 and 13 (Table 3) do, in fact, constitute a single gene (Orr *et al*., 1997).

Nine other ORFs appear to be unique to *T. pyriformis* mtDNA, although it is notable that *P. aurelia* mtDNA contains ORFs of similar size and in the same position as five of these unique Tetrahymena

Table 2. Protein-coding genes in mitochondrial DNA

	Tetrahymena pyriformis	*Paramecium aurelia*	*Acanthamoeba castellanii*	*Marchantia polymorpha*
nad1	■	■	■	■
nad2	■	■	■	■
nad3	■	■	■	■
nad4	■	■	■	■
nad4L	○	○	■	■
nad5	■	■	■	■
nad6	○	○	■	■
nad7	■	■[a,b]	■	□
nad9	■	■[a,c]	■	■
nad10	■	■[a,d]	○	○
nad11	○	○	■	○
cob	■	■	■	■
cox1	■	■[a,e]	■	■
cox2	■	■[a,f]	■	■
cox3	○	○	■	■
atp1	○	○	■	■
atp6	○	○	■	■
atp9	■	■	■	■
yejR (*ccl1*)	■	■	○	■[j]
rps1	○	○	○	■
rps2	○	○	■	■
rps3	■	■[a,g]	■	■
rps4	○	○	■	■
rps7	○	○	■	■
rps8	○	○	■	■
rps10	○	○	○	■
rps11	○	○	■	■
rps12	■	■	■	■
rps13	■	■[a,h]	■	■
rps14	■	■	■	■
rps19	■	○	■	■
rpl2	■	■	■	■
rpl5	○	○	■	■
rpl6	○	○	■	■
rpl11	○	○	■	○
rpl14	■	■	■	○
rpl16	■	■[a,i]	■	■

T. pyriformis (this work); *P. aurelia* (Pritchard *et al*., 1990b); *A. castellanii* (Burger *et al*., 1995); *M. polymorpha* (Oda *et al*., 1992). Genes underlined are also found in animal mtDNA.

Symbols are: ■, gene present; ○, gene absent; □, pseudogene.

[a] See Pritchard *et al*. (1990b).
[b] ORF400.
[c] P1.
[d] *psbG*.
[e] Includes ORF11.
[f] Includes ORF15.
[g] ORF7.
[h] ORF8.
[i] ORF26.
[j] ORF509 (*ymf4*); Oda *et al*. (1992).

ORFs (Table 4). However, we cannot convincingly demonstrate that these positionally equivalent ORFs are homologous. Of the unidentified ORFs in *T. pyriformis* mtDNA, the most notable is *ymf77*, which has no counterpart in *P. aurelia* mtDNA and whose size (1386 codons) makes it the longest ORF in this genome. The length of *ymf77* is remarkable, considering the high A + T content of the DNA in which it resides, which would lead one to expect a high density of potential translation termination codons (particularly TAA) in this region.

Major transcriptional clusters are oriented toward the leftward and rightward ends of *T. pyriformis* mtDNA, starting approximately in the middle of the genome (Figure 1). A cluster of three genes (*nad7-rps14-orf178*) of opposite transcriptional orientation divides the leftward transcriptional unit into two roughly equal parts (Figure 1). As observed in other completely sequenced protist mtDNAs (e.g. *Acanthamoeba castellanii*, Burger *et al*., 1995; *Reclinomonas americana*, Lang *et al*., 1997; see also Gray *et al*., 1998), genes are tightly packed in *T. pyriformis* mtDNA, with intergenic spacers (ranging in size from 0-436 bp) constituting only 4.0 % of the total sequence, and having a higher A + T content (87.5 %, excluding telomeres) than coding regions (78.3 %). However, apparent overlapping of genes, which is prominent in some other mtDNAs (e.g. that of *A. castellanii*; Burger *et al*.,

Table 3. Ciliate-specific ORFs in *Tetrahymena pyriformis* and *Paramecium aurelia* mtDNAs

ymf[a]	ORF, *T. pyriformis*[b]	ORF, *P. aurelia*[b]	Pritchard *et al.* (1990b)	Optimized *z* value (RDF2)[c]	BLASTP score[c]	BLASTP probability[c]
ymf56	*orf97*	*orf78*		3.55	56	1.7e−5
ymf57	*orf100*	*orf100*		11.38	219	2.4e−31
ymf58	*orf116*	*orf113*	ORF1	10.00	162	1.3e−22
ymf59	*orf152*	*orf124*		5.23	65	3.7e−6
ymf60	*orf178*	*orf178*		17.18	111	1.2e−22
ymf61	*orf237*	*orf189*		7.35	57	1.1e−8
ymf62	*orf256*	*orf265_2*		14.11	213	5.9e−50
ymf63	*orf276*	*orf314*		6.34	118	1.1e−11
ymf64	*orf328*	*orf234*		7.98	167	2.9e−19
ymf65	*orf365*	352[j] { *orf196*[d]	ORF3	6.21	110	2.5e−14
		352[j] { *orf156*[e]	ORF4	15.42	295	8.1e−41
ymf66	*orf448*	410[j] { *orf189*[f]	ORF16	4.35	72	8.9e−8
		410[j] { *orf221*[g]	ORF17	22.41	207	7.0e−26
ymf67	*orf443*	355[j] { *orf265_1*[h]	ORF12	−0.19	n.s.[k]	n.s.[k]
		355[j] { *orf90*[i]	ORF13	12.32	180	3.2e−21
ymf68	*orf593*	*orf393*	ORF14	13.98	288	7.3e−56

These ORFs are all in the same relative positions with respect to flanking genes in the two ciliate mitochondrial genomes.

[a] The *ymf* designations for ORFs follow the recommendations of the Commission on Plant Gene Nomenclature (1993). The particular designations used here have been registered with the Commission.

[b] Numbers indicate number of amino acid residues in each *orf*.

[c] Threshold values for homology are considered to be an RDF2 *z* value >5 and/or a BLASTP score >50 with a probability <1.0e−2. These values were chosen based on simulations with homologous pairs of identified *T. pyriformis* and *P. aurelia* mitochondrial genes; in these cases, the corresponding values ranged from a low of 3.63, 64 and 1.3e−2 (*rps14*) to a high of 115.40, 532 and 4.6e−115 (*cox1*). In the case of RDF2, the *z* value is calculated by subtracting the mean score of randomly shuffled sequences from the score of the unshuffled sequence and then dividing by the standard deviation of the distribution of shuffled scores. Pearson (1990) suggests that ''one should be skeptical of conclusions based on sequence similarity scores with *z* values less than 3, and more confident when *z* values are greater than 6.''

[d] *ymf65_a*.

[e] *ymf65_b*.

[f] *ymf66_a*.

[g] *ymf66_b*.

[h] *ymf67_a*; separate ORF in the original report (Pritchard *et al.*, 1990b) but part of a single continuous ORF with *ymf67_b* in Orr *et al.* (1997).

[i] *ymf67_b*; see footnote h.

[j] Total length of adjacent ORFs.

[k] Not statistically significant.

1995), is rare in the *T. pyriformis* mitochondrial genome.

Structure and sequence of highly divergent mitochondrial protein-coding genes

In *T. pyriformis* and *P. aurelia*, mtDNA-encoded proteins display unusual sequence characteristics that are not seen in the same proteins in other eukaryotes. Multiple protein alignments revealed that the Nad2 protein of both ciliates lacks the otherwise conserved N-terminal ~350 residues of a typical Nad2, and sequences encoding this region cannot be identified elsewhere in either mitochondrial genome. Similarly, the expected C-terminal ~100 residues of Nad5 are missing in the ciliate versions of this protein, whereas both ciliate sequences feature prominent in-frame N-terminal extensions (149 residues in Tetrahymena, 72 residues in Paramecium) before the first Met. These extensions, if present in the mature protein, would represent additional sequence compared to Nad5 proteins from other organisms.

In addition, conspicuous large inserts are uniquely present in the deduced Cox1 and Cox2 protein sequences of the two ciliates; these inserts are of the order of 100 and 300 amino acid residues, respectively, and are located after positions corresponding to residues 124 (Cox1) and 159 (Cox2) in the *Marchantia polymorpha* homologs. The two ciliate Cox1 inserts share sequence similarity with one another, as do the two Cox2 inserts. In *P. aurelia*, the Cox1 and Cox2 inserts were annotated as in-frame upstream ORFs (ORF15 and ORF11, respectively) (Pritchard *et al.*, 1990b), because the N-terminal portions of these two genes were not recognized due to the extreme divergence of the ciliate sequences. Indeed, within the N-terminal region, the Tetrahymena and Paramecium Cox2 proteins share only limited similarity with the homologous protein in other organisms, the most obvious signature being a conserved motif, QWYW, located at positions 121-124 in the *T. pyriformis* Cox2 sequence.

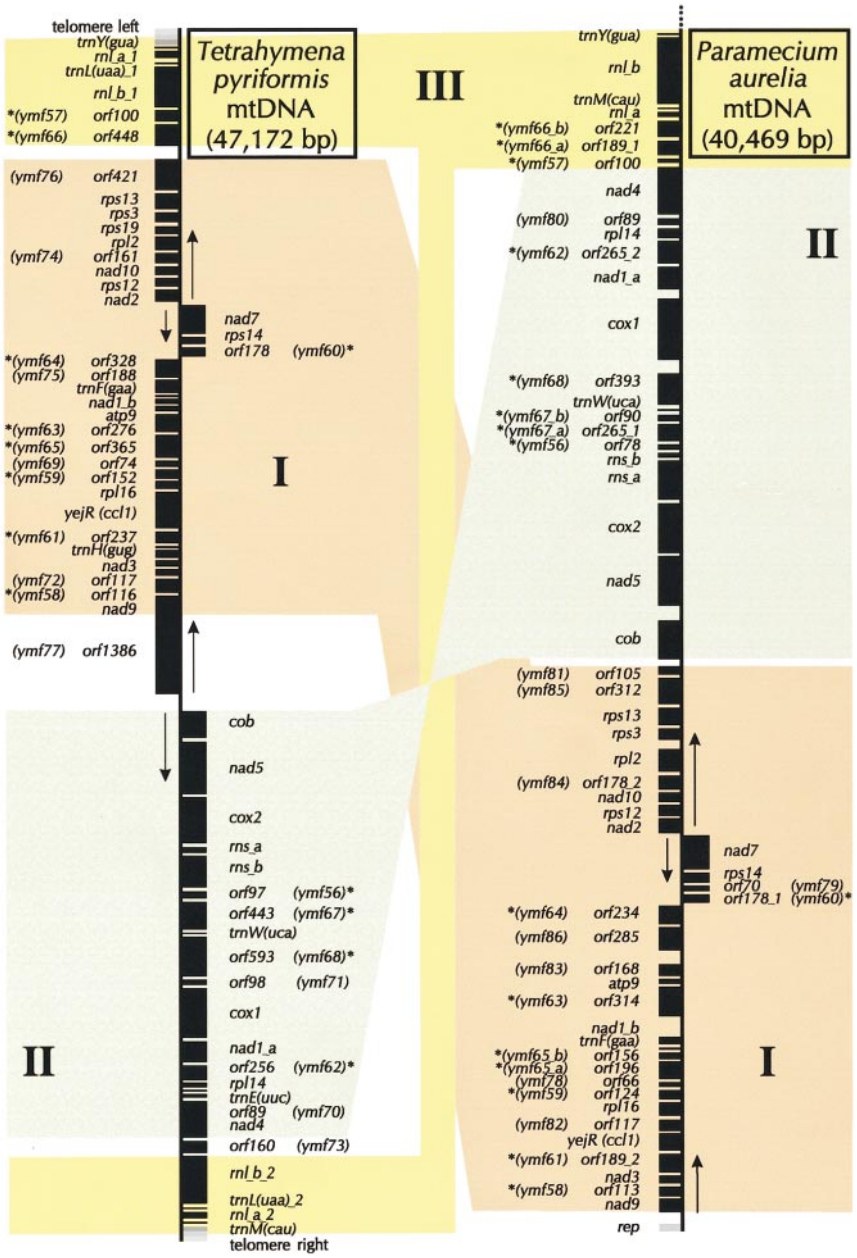

Figure 1. Gene maps of *T. pyriformis* and *P. aurelia* mtDNAs (see Tables 1-4 for additional information on genes in Tetrahymena mtDNA). The mitochondrial genome size listed for *T. pyriformis* excludes telomere repeats. The ''left'' and ''right'' ends of the Tetrahymena mitochondrial genome are designated with reference to the subterminal inverted repeats that contain *rnl*, as in previous reports (Heinonen *et al.*, 1987, 1990). Black boxes, genes and ORFs; grey boxes, repeat arrays. Blocks of genes/ORFs positioned to the left and right of the Paramecium and Tetrahymena ''backbones'' are transcribed from top to bottom and bottom to top, respectively, as indicated by the arrows. Duplicated genes are distinguished by the designations *_1* and *_2*, as are non-homologous ORFs of the same size. The separate portions of fragmented genes (*rns*, *rnl*, *nad1*) are denoted by *_a* and *_b*. *ymf* assignments are given in parentheses (see Tables 3 and 4), with *_a* and *_b* in the Paramecium map designating individual ORFs that are likely part of a single gene, as discussed in the text.

Finally, the deduced ciliate Rps13 sequence features a long C-terminal extension of 100 (Paramecium) or 135 (Tetrahymena) amino acid residues. These ciliate-specific extensions share limited sequence similarity.

Codon usage in ciliate mitochondria

As discussed in the accompanying paper (Edqvist *et al.*, 2000), protein sequence alignments and transcript mapping allowed us to infer the

Table 4. Unique ORFs in *Tetrahymena pyriformis* and *Paramecium aurelia* mtDNAs

T. pyriformis			*P. aurelia*	
ORF[b]	Assigned *ymf*	Positional equivalence[a]	ORF[b]	Assigned *ymf*
orf74	*ymf69*	X	*orf66*	*ymf78*
			orf70	*ymf79*
orf89	*ymf70*	X	*orf89*	*ymf80*
orf98	*ymf71*			
orf117	*ymf72*			
orf160	*ymf73*			
			orf117	*ymf82*
orf161	*ymf74*	X	*orf178*	*ymf84*
orf188	*ymf75*	X	*orf168*	*ymf83*
orf421	*ymf76*	X	*orf105*	*ymf81*
			orf312	*ymf85*
orf1386	*ymf77*			
			orf285	*ymf86*

[a] Pairs of ORFs that are positionally equivalent in Tetrahymena and Paramecium mtDNAs are shown on the same line and marked by X. Note that Tetrahymena *ymf76* corresponds to two positionally adjacent ORFs, *ymf81* and *ymf85*, in Paramecium mtDNA.
[b] ORF number denotes number of amino acid residues.

translation initiation codons that are used in *T. pyriformis* mitochondria. In addition to the standard ATG codon, start codons include variants of ATG that differ in either the first or third position (ATA, ATT, GTG, TTG).

Codon usage was analyzed separately for three classes of protein-coding sequences in *T. pyriformis* mtDNA: identified genes, ciliate-specific ORFs and unique ORFs (Table 5). In all three cases, only TAA is used to terminate translation, and three of the six Arg codons (CGC, CGA, CGG) do not appear in any of these genes. The Gly codon GGC is absent from both the ciliate-specific and unique gene sets and is rarely used in the collection of identified genes (12 out of 288 GGN codons). As expected for an extremely A + T-rich genome, codons ending in A or T vastly outnumber the synonymous codons ending in G or C. As in many other mitochondrial translation systems (Gray *et al*., 1998), TGA is used in addition to the standard TGG codon to specify Trp in *T. pyriformis* mitochondria; in fact, TGA is used preferentially, accounting for >90 % of all Trp codons in each of the three protein-coding classes in *T. pyriformis* mtDNA (Table 5).

To a striking degree, codon usage in the ciliate-specific and unique ORF classes parallels that in identified protein-coding genes (Table 5). A good example of this correspondence is the six Ser codons, whose proportional use is very similar in the three gene classes. Absence of certain codons, virtually exclusive use of others, and overall frequency of occurrence of codons specifying the same amino acid all strongly favor the conclusion

Table 5. Codon usage in protein-coding genes in *T. pyriformis* mtDNA

TTT	F	88*(90, 90)*	**TCT**	S	34*(35,31)*	**TAT**	Y	81*(83, 84)*	**TGT**	C	87*(94,74)*
TTC	F	12*(10, 10)*	**TCC**	S	4*(4, 4)*	**TAC**	Y	19*(17, 16)*	**TGC**	C	13*(6,26)*
TTA	L	*77(76, 67)*	**TCA**	S	33*(31,32)*	**TAA**	*	100*(100,100)*	**TGA**	W	98*(97,92)*
TTG	L	3*(2, 5)*	**TCG**	S	3*(2, 3)*	**TAG**	*	*-(-, -)*	**TGG**	W	2*(3, 8)*
CTT	L	6*(7, 10)*	**CCT**	P	57*(49,49)*	**CAT**	H	70*(68, 65)*	**CGT**	R	3*(3, 1)*
CTC	L	1*(0, 1)*	**CCC**	P	3*(4, 5)*	**CAC**	H	30*(32, 35)*	**CGC**	R	*-(-, -)*
CTA	L	12*(14, 15)*	**CCA**	P	36*(41,30)*	**CAA**	Q	94*(95, 96)*	**CGA**	R	*-(-, -)*
CTG	L	0*(1, 2)*	**CCG**	P	5*(6,16)*	**CAG**	Q	6*(5, 4)*	**CGG**	R	*-(-, -)*
ATT	I	34*(32, 31)*	**ACT**	T	39*(42,42)*	**AAT**	N	78*(81, 79)*	**AGT**	S	22*(22,21)*
ATC	I	6*(6, 4)*	**ACC**	T	6*(6, 7)*	**AAC**	N	22*(19, 21)*	**AGC**	S	4*(7, 9)*
ATA	I	60*(61, 65)*	**ACA**	T	55*(51,48)*	**AAA**	K	97*(97, 95)*	**AGA**	R	95*(95,98)*
ATG	M	100*(100,100)*	**ACG**	T	0*(1, 3)*	**AAG**	K	3*(3, 5)*	**AGG**	R	3*(2, 1)*
GTT	V	42*(46, 32)*	**GCT**	A	61*(58,35)*	**GAT**	D	81*(78, 83)*	**GGT**	G	*77(82,74)*
GTC	V	6*(2, 9)*	**GCC**	A	4*(9,20)*	**GAC**	D	19*(22, 17)*	**GGC**	G	4*(-, -)*
GTA	V	48*(50, 52)*	**GCA**	A	31*(30,42)*	**GAA**	E	91*(92, 94)*	**GGA**	G	15*(16,23)*
GTG	V	3*(2, 7)*	**GCG**	A	4*(4, 2)*	**GAG**	E	9*(8, 6)*	**GGG**	G	3*(2, 3)*

Codons are shown in bold lettering together with the one-letter designations for the corresponding amino acids. Numbers indicate the percentage use of each codon for a given amino acid. The first number listed refers to identified protein-coding genes (Table 1), whereas the numbers shown in parentheses and italics refer to ciliate-specific (Table 2) and unique (Table 3) gene classes, respectively. The total number of codons evaluated was 6322 in 22 identified ORFs, 3602 in 13 ciliate-specific ORFs, and 2703 in nine unique ORFs. (−), complete absence of the codon in question from the entire set of genes in a particular class; (*), a termination codon.

that the ciliate-specific and unique ORFs are, in fact, functional protein-coding genes.

On the other hand, consideration of inferred amino acid sequence leads us to question the authenticity of *ymf71* and *ymf73*, two of the four unique Tetrahymena ORFs that have no positional homologs in Paramecium mtDNA (i.e. *ymf71*, *ymf72*, *ymf73* and *ymf77*) (Table 4). Both *ymf71* and *ymf73* have an A + T content that is appreciably higher than the average for coding regions (91.2 % for *ymf71* and 85.3 % for *ymf73*, compared with an average of 78.3 %); in both instances, codons composed only of A and T predominate, with codon frequencies closely approximating a random distribution expected from the A + T content of the ORF; and both ORFs completely lack any Gly or Gln residues, with *ymf71* also lacking Pro, Ala, Cys and Arg. These peculiarities suggest that *ymf73* and especially *ymf71* may be fortuitous ORFs rather than real genes.

A distinctly different codon usage pattern is seen in *P. aurelia* mtDNA, whose coding regions are considerably less A + T-rich than in the case of *T. pyriformis* mtDNA (58.1 % *versus* 78.3 %). Except for CCG (Pro), which is not found in the ciliate-specific ORFs, all codons are used in all three classes of *P. aurelia* mtDNA, including TAG as well as TAA as a termination codon. Synonymous codons are employed in a much more balanced fashion, with TGA and TGG appearing with almost equal frequency; moreover, use of codons ending in C or G often exceeds that of synonymous codons ending in T or A. However, as in the case of *T. pyriformis* mtDNA, overall codon frequency is very similar in the three classes of mitochondrial protein-coding sequences in *P. aurelia*.

Transfer RNA genes

T. pyriformis mtDNA encodes only eight tRNA genes corresponding to seven distinct tRNA species that together recognize only 13 different codons (Table 1). These genes account for all of the mtDNA-encoded tRNA sequences previously identified in this organism at either the DNA (Suyama, 1985; Heinonen *et al*., 1987; Suyama *et al*., 1987; Suyama & Jenney, 1989) or RNA (Schnare *et al*., 1985, 1995) level. Suyama (1986) had inferred, on the basis of mitochondrial tRNA fractionation and hybridization experiments, that *T. pyriformis* mtDNA encodes fewer than ten tRNA species. Our results formally confirm the long-standing proposal (Chu *et al*., 1975; Suyama, 1982) that nucleus-encoded tRNAs must be imported from the cytosol into *T. pyriformis* mitochondria, to supplement the mtDNA-encoded tRNA species, whose number is not sufficient to support mitochondrial translation. *P. aurelia* mtDNA encodes even fewer tRNAs than *T. pyriformis* mtDNA, additionally lacking *trnL*, *trnH* and *trnE* (Pritchard *et al*., 1990b).

As expected, the $tRNA^{Trp}$ sequence predicts a UCA anticodon, able to decode both UGG and UGA codons. In fact, the corresponding mitochondrial tRNA from *T. thermophila* has been shown to function as a UGA suppressor tRNA in an *in vitro* translation assay (Hekele & Beier, 1991). With the exception of the $tRNA^{Phe}$ (Suyama, 1985; Schnare *et al*., 1985) and $tRNA^{Met}$ (Heinonen *et al*., 1987; Schnare *et al*., 1995), the other mtDNA-encoded tRNAs have conventional secondary structures. The unusual structural features of the $tRNA^{Met}$, which include a truncated D stem and loop and two adjacent pseudouridine residues in the anticodon loop (Schnare *et al*., 1995), may be implicated in the recognition of variant translation initiation codons (see Edqvist *et al*., 2000). In spite of these peculiarities, this structurally deviant tRNA has the potential to assume the L-shaped tertiary structure characteristic of a normal tRNA (Steinberg & Cedergren, 1994). *P. aurelia* mtDNA also encodes a homolog of this unusual $tRNA^{Met}$ (Heinonen *et al*., 1987).

Sequence polymorphisms in the duplicated *rnl* gene

In *T. pyriformis* mitochondria, the LSU rRNA comprises two separate species, LSUα (encoded by *rnl_a* and corresponding to the first 280 nt of the LSU rRNA sequence) and LSUβ (encoded by *rnl_b* and representing the rest of the sequence; Heinonen *et al*., 1987). We had previously shown that the duplicate *rnl* genes in *T. pyriformis* mtDNA are not identical in sequence but differ at five positions, one (di6, Table 6) within *rnl_a* and the remaining four (di2 to di5, Table 6) within *rnl_b*

Table 6. Sequence heterogeneity in the mitochondrial *rnl* genes and their rRNA products in *T. pyriformis*

	rnl_a/LSUα	*rnl_b*/LSUβ				
DNA/RNA preparation	di6 (182)	di1 (554)	di2 (803)	di3 (1907)	di4 (1911)	di5 (2074)
DNA-1[a], left cistron	G	C	G	T	C	C
DNA-1[a], right cistron	A	C	A	A	T	T
RNA-1[a]	**A > G**	**C**	**A > G**	**A > U**	**U > C**	**U > C**
DNA-2[b], left cistron	G	C	G	T	C	T
DNA-2[b], right cistron	A	T	A	T	C	T
RNA-2[b]	**A > G**	**U > C**	**A > G**	**U**	**C**	**U**

di1 to d6 represent polymorphic sites (see annotation accompanying GenBank acc. no. AF160864), with the number in parentheses indicating the position of that site in the mature LSUα or LSUβ rRNA (see Heinonen *et al*., 1990).

[a] Heinonen *et al*. (1987, 1990).

[b] This work.

(Heinonen *et al*., 1990). Here, clones representing these polymorphic sites were assigned to the left or right ends of the genome by linkage with one another and with unique flanking markers that are left-end or right-end-specific. In this analysis, three of the previously polymorphic sites (di3 to di5, Table 6) were found to be homogeneous. On the other hand, a new polymorphic site (di1, Table 6), not seen in the previous study, was noted.

Because sequencing data from the two studies are unambiguous, it appeared that these discrepancies might represent real differences in the sequence of the two mtDNA preparations used in these analyses. These samples were prepared some eight years apart, from separate isolates of the same strain of *T. pyriformis* ST (both isolates originating in the laboratory of Y. Suyama). In the previous study, one of the polymorphic differences was manifested by sequence heterogeneity in the corresponding RNA species (Heinonen *et al*., 1987); therefore, we used a reverse transcriptase sequencing approach to evaluate the polymorphic positions in the rRNA that had been prepared at the same time as the respective mtDNA preparations. Representative sequencing gels (Figure 2) document the di6 polymorphism (see Table 6) shared between the two DNA preparations (at LSU rRNA position 182; Figure 2(a)), as well as the di1 polymorphism (Table 6) that is seen only in the more recently isolated mtDNA preparation (corresponding to position 554 in the LSU rRNA; Figure 2(b) and (c)). The results of rRNA sequencing (Table 6) fully confirmed the polymorphisms and verified the differences detected in each of the independently determined mtDNA sequences within the *rnl* regions. Thus, in the interval between different mtDNA preparations, there appears to have been homogenization at three previously polymorphic sites and the appearance of a new one. This new site (correponding to position 554 in the LSU rRNA sequence) is localized in a single-stranded region of the secondary structure (Heinonen *et al*., 1990), with the alternative nucleotides at this position (C or U) having no obvious effect on structure.

The RNA sequencing results (see, e.g. Figure 2(a)) also provided evidence that ''rightward'' LSUα and LSUβ transcripts predominate in the steady-state RNA preparation (Table 6), confirming previous results with the LSUα rRNA (Heinonen *et al*., 1987).

Comparison of gene organization in *T. pyriformis* and *P. aurelia* mitochondrial genomes: evolutionary rearrangement in ciliate mtDNAs

Assuming that the positionally equivalent ORFs listed in Table 4 are highly divergent homologs, there are very few differences in the genetic information contents of *T. pyriformis* and *P. aurelia* mtDNAs. Moreover, large blocks of genes are colinear in the two genomes, although the blocks themselves have been rearranged relative to one another. One gene array (block I), comprising about one-third of the genome, extends from *nad9* to *ymf76* in *T. pyriformis* mtDNA; except for a small segment from *nad7* to *ymf60*, all of the genes in

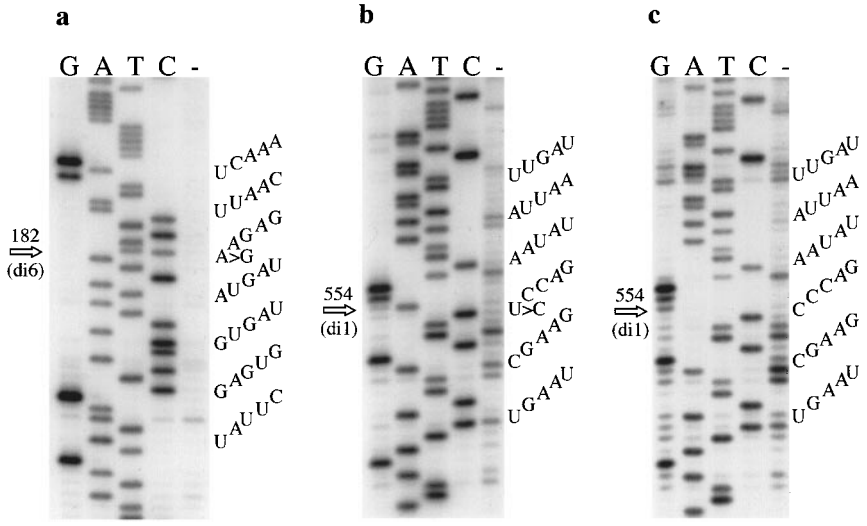

Figure 2. Oligonucleotide-primed RT sequencing of *T. pyriformis* mitochondrial LSU rRNA. Lanes G, A, T, C indicate the dideoxynucleotide incorporated during RT sequencing. The control lanes (−) represent primer extension products generated in the absence of dideoxynucleotides. Total RNA (prepared ''new'' for this study) was used as template to determine portions of the (a) LSUα and (b) LSUβ rRNA sequences (see the text). The LSUβ rRNA that had been gel-purified from an ''old'' RNA preparation (Heinonen *et al*., 1987) and stored in 50 % ethanol at −20 °C for several years was also used as a (c) template for RT sequencing. Deduced RNA sequences are shown to the right of the autoradiograms. The arrows indicate positions at which sequence heterogeneity has been identified (di1 and di6 represent sequence polymorphisms listed in Table 6).

block I are in the same transcriptional orientation (Figure 1). Differences in block I between the two mitochondrial genomes involve eight genes: absence of *rps19*, *trnH* and *ymf72* in *P. aurelia* mtDNA, absence of *ymf79, ymf86* and *ymf82* in *T. pyriformis* mtDNA, and transposition of the *trnF* and *nad1_b* genes relative to *ymf63* and *atp9*. Although *trnF* and *nad1_b* are adjacent to one another in both genomes, they are located immediately after *ymf65* in *P. aurelia* mtDNA, but two genes further downstream (after *atp9*) in *T. pyriformis*. Moreover, although the transcriptional orientation of the two genes is maintained with respect to their neighbors in block I, their positions relative to one another are switched, i.e. *trnF* → *nad1_b* in Paramecium *versus nad1_b* → *trnF* in Tetrahymena.

Block II, comprising another third of the genome, extends from *cob* to *nad4*. In this block also, all genes are transcribed in the same direction, with differences limited to absence of *ymf71* and *trnE* from *P. aurelia* mtDNA. In the time since these two ciliates diverged from a common ancestor, block II has been inverted relative to block I in one of the two genomes (Figure 1). In *T. pyriformis* mtDNA, a unique ORF, *ymf77*, is located between the two blocks.

A major difference between the two mtDNAs is seen in the *rnl* gene region (block III), which is duplicated and rearranged in *T. pyriformis* relative to *P. aurelia*. As documented, *rnl* is divided into two separate coding regions (*rnl_a* and *rnl_b*) in both *T. pyriformis* and *P. aurelia* mtDNAs; in the former organism, *rnl_a* and an immediately downstream *trnM* gene have apparently been transposed to the end of *rnl_b*, with a *trnL* gene (which *P. aurelia* mtDNA lacks) interposed between *rnl_b* and *rnl_a* (Heinonen *et al*., 1987) (Figure 1). In *T. pyriformis*, a *trnY* gene replaces *trnM* in one of the two copies of the *rnl* repeat; in *P. aurelia*, *trnM* and *trnY* genes flank the single-copy *rnl_b* cistron (see Figure 1). Although *ymf57* and *ymf66* are maintained in block III in Paramecium mtDNA and in one of the duplicated regions of block III in *T. pyriformis* mtDNA, the transcriptional orientation of the two genes is switched in the one genome relative to the other. In *T. pyriformis* mtDNA, the duplicated *rnl* region is located in a subterminal inverted repeat, with 31 bp telomeric sequences tandemly repeated at the ends of the linear chromosome. In *P. aurelia* mtDNA, the *rnl* gene region is positioned at one end of the mitochondrial genome whereas a telomere-like replication region (rep), consisting of a tandem array of 35 bp direct repeats, is found at the other end of the linear chromosome. Because an estimated 200 bp remains unsequenced at the *rnl* end of *P. aurelia* mtDNA (Pritchard *et al*., 1990b), it is possible that a telomeric terminal structure also exists in this linear mitochondrial genome.

Phylogenetic analysis

The highly divergent nature of ciliate mitochondrial protein-coding sequences (Pritchard *et al*., 1990a; Cummings, 1992) is manifested in very long branches in phylogenetic trees based on a concatenated set of mitochondrial proteins (Figure 3). In such a tree, the Paramecium and Tetrahymena sequences form a monophyletic grouping. The long branches suggest a greatly accelerated rate of mtDNA sequence change in the lineage leading to the two ciliates, such that the common ancestor of Paramecium and Tetrahymena must already have possessed a highly divergent mitochondrial genome. Whether an accelerated rate of sequence evolution is characteristic of the ciliate phylum as a whole remains to be investigated. A long branch leading to ciliates is also evident in phylogenetic trees based on mitochondrial rRNA sequences (e.g. Gray & Spencer, 1996). In contrast, in trees based on nucleus-encoded Hsp70 sequences (Budin & Philippe, 1998), branch lengths for the ciliate sequences are not appreciably greater than for those of other eukaryotes. This apparent acceleration in the rate of divergence of ciliate mtDNA-encoded sequences (particulary protein-coding ones) is an evolutionary problem whose basis remains obscure at this time.

The monophyly of ciliates, and a common evolutionary origin of apicomplexans, dinoflagellates and ciliates, is well supported by phylogenies based on nuclear rRNA sequences (e.g. Sogin 1991, 1997). The common ancestry of ciliates and apicomplexans is also supported by more recent analyses using a nucleus-encoded Hsp70 data set (Budin & Philippe, 1998). In the tree shown in Figure 3, an affiliation between Plasmodium (apicomplexan) and ciliates is strongly supported (100% bootstrap); however, these results have to be interpreted with caution, because the rapidly evolving Plasmodium and ciliate branches may be artifactually grouped together due to long-branch attraction (Felsenstein, 1988).

The phylogenetic relationship between slime molds and alveolates has been the subject of numerous studies whose conclusions have not always been entirely consistent. Phylogenetic trees based on mitochondrial protein sequence data (Figure 3) are in agreement with previous conclusions that ciliates plus apicomplexans share a common ancestor (Gajadhar *et al*., 1991), as do dictyostelid plus acellular slime molds (Baldauf & Doolittle, 1997). Additionally, however, in the mitochondrial tree shown in Figure 3, and in contrast to what is seen in nuclear gene trees, these two clades form a single lineage that originates from an unresolved radiation point (the ''crown radiation'') within the eukaryotic lineage. This affiliation of ciliates/apicomplexa with slime molds is robustly supported by bootstrap analysis, although again one must caution against the possibility of long-branch-attract artifacts. When the Plasmodium and two ciliate sequences are

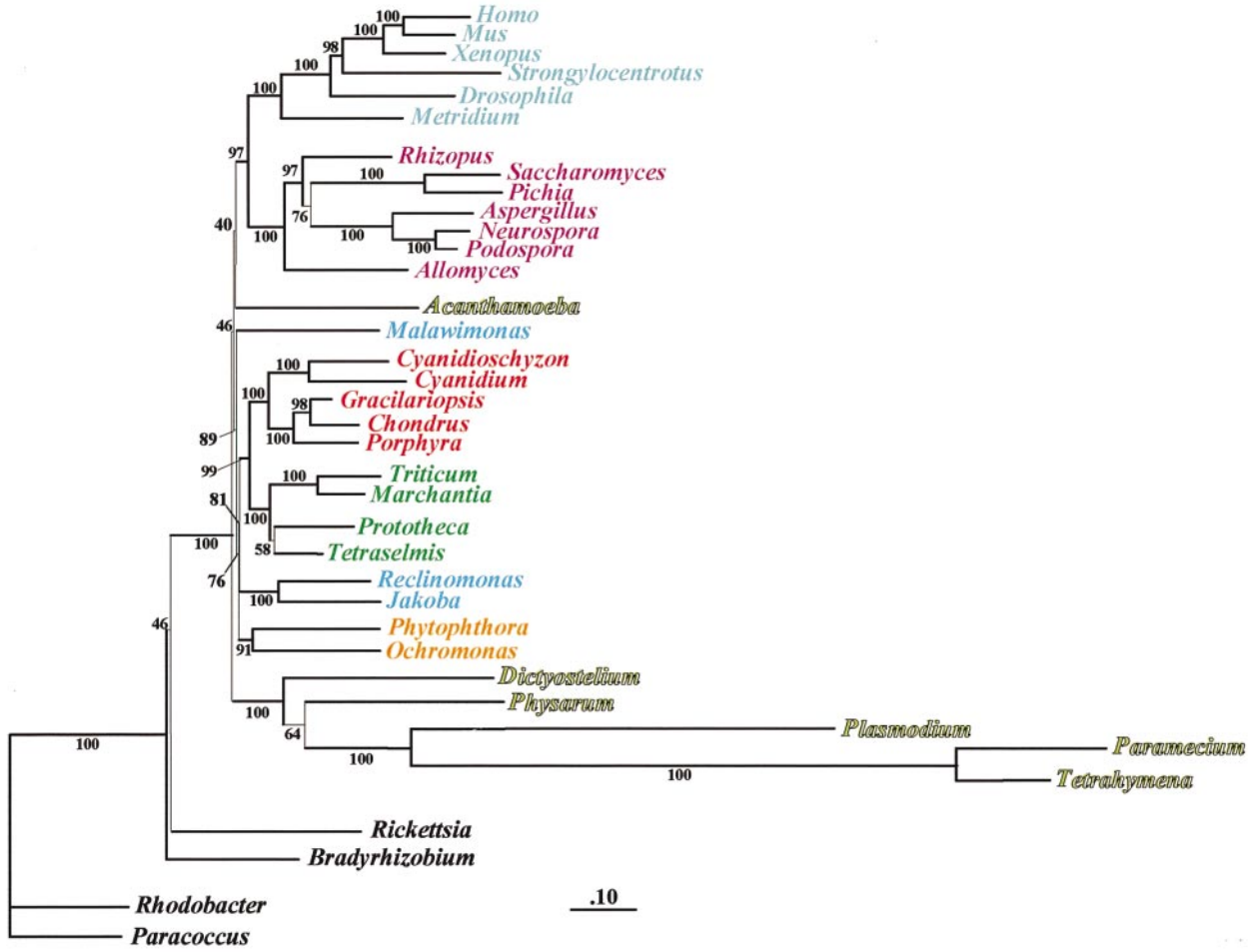

Figure 3. Phylogenetic tree inferred from mtDNA-encoded protein sequences. An alignment of the concatenated amino acid sequences of Cob, Cox1, Cox2 and Cox3 was used except for species where only a subset of these proteins are mitochondrially encoded (Tetrahymena, Paramecium, Plasmodium), or where only incomplete mtDNA sequences are currently available (Physarum, Tetraselmis, Gracilariopsis). Only unambigously aligned portions of these protein sequences were included in the analysis. The tree shown was inferred using the most recent implementation of PROTDIST/FITCH (Phylip, version 3.6; Felsenstein, 1993), which allows a Jin/Nei correction for unequal rates of change at different amino acid positions. The variation coefficient used was 0.3. Bootstrap support (%) was calculated from 1000 replicates using the PARBOOT parallel bootstrapping package (http://megasun.bch.umontreal.ca/ogmp/ogmpid.html). The scale bar (.10) denotes mean number of substitutions per site. Except in the case of poorly supported branches (thin lines), the same tree topology was obtained when maximum likelihood approaches (PROTML, PUZZLE) were used. Animals (light blue), fungi (purple), red algae (red), green algae/plants (green), jakobids (blue), ciliates/apicomplexans/slime molds/rhizopods (olive), stramenopiles (orange) and bacteria (black). Species (with GenBank accession numbers in parentheses) are: Homo, *H. sapiens* (chordate; J01415); Mus, *M. musculus* (chordate; J01420) Xenopus, *X. laevis* (chordate; (M10217); Strongylocentrotus, *S. purpuratus* (echinoderm; X12631); Drosophila, *D. yakuba* (arthropod; X03240); Metridium, *M. senile* (cnidarian; AF000023). Rhizopus, *R. stolonifer* (chytridiomycete fungus; Paquin *et al.*, 1997); Saccharomyces, *S. cerevisiae* (ascomycete fungus; P00163, V00694, J05007, J01478); Pichia, *P. candensis* (*Hansenula wingei*) (ascomycete fungus; D31785); Aspergillus, *A.* (*Emericella*) *nidulans* (ascomycete fungus; J01387, X15441, X06960); Neurospora, *N. crassa* (ascomycete fungus; K01181, X01850, K00825, V00668); Podospora, *P. anserina* (ascomycete fungus; X55026); Allomyces, *A. macrogynus* (chytridiomycete fungus; U41288); Acanthamoeba, *A. castellanii* (rhizopod; U12386); Malawimonas, *M. jakobiformis* (jakobid; Burger *et al.*, OGMP); Cyanidioschyzon, *C. merolae* (rhodophyte; D89861); Cyanidium, *C. caldarium* (rhodophyte; Z48930); Chondrus, *C. crispus* (rhodophyte; Z47547); Porphyra, *P. purpurea* (rhodophyte; AF114794); Gracilariopsis, *G. lemaneiformis*, Cox1,2,3 (rhodophyte; AF118119); Triticum, *T. aestivum* (angiosperm; P07747 (protein), Y00417, X01108, P15953 (protein)); Marchantia, *M. polymorpha* (bryophyte; M68929); Prototheca, *P. wickerhamii* (chlorophyte; U02970); Tetraselmis (*Platymonas*) *subcordiformis* (chlorophyte; Z47797); Reclinomonas, *R. americana* (jakobid; AF007261); Jakoba, *J. libera* (jakobid; Burger *et al.*, OGMP); Phytophthora, *P. infestans* (oomycete; U17009); Ochromonas, *O. danica* (chrysophyte; G. Burger & A. Coleman, unpublished results); Dictyostelium, *D. discoideum* (dictyostelid slime mold; AB000109); Physarum, *P. polycephalum* (slime mold; AF084526, L14769, AF079799); Plasmodium, *P. falciparum* (alveolate; M76611); Paramecium, *P. aurelia* (ciliate; X15917); Tetrahymena; *T. pyriformis* (ciliate; AF160864, this work); Rickettsia, *R. prowazekii* (α-Proteobacterium; AJ235270 to 73); Bradyrhizobium, *B. japonicum* (α-Proteobacterium; J03176, X68547); Rhodobacter, *R. sphaeroides* (α-Proteobacterium; X56157, X62645, M57680, C45164 (PIR)); Paracoccus, *P. denitrificans* (α-Proteobacterium; X05829, M17522, X05934, X05828). (All unpublished protein sequences used in this analysis are available at http://megasun.bch.umontreal.ca/People/lang/FMGP/proteins.html).

removed from the current data set, the tree topology shown in Figure 3 was unchanged except that Acanthamoeba moved from its ill-defined divergence within the crown radiation to a divergence point basal to that of the two slime molds, Physarum and Dictyostelium. Although this particular result is not well supported by bootstrap analysis, other mitochondrial data do suggest a specific phylogenetic relationship between Acanthamoeba and Dictyostelium. For example, in the mitochondrial genome of these two amoebae, *cox1* and *cox2* are joined in-frame to generate a single ORF that encodes both Cox1 and Cox2 proteins; moreover, both genomes encode a limited and almost identical set of tRNA genes (see Gray *et al.*, 1998). We argue that these unique shared characters, especially the *cox1-cox2* gene fusion, provide strong support for a specific phylogenetic affiliation between Acanthamoeba and Dictyostelium. Finally, in contrast to what is seen with nuclear rRNA data, trees based on actin, β-tubulin and elongation factor alpha unite slime molds in a single coherent (monophyletic) group, in close relationship to the animal-fungal clade (Baldauf & Doolittle, 1997). Sequences from additional ciliates, rhizopod amoebae and slime molds will be necessary to further explore a possible evolutionary link between these protist groups at the level of the mitochondrial genome.

Discussion

Determination of the complete sequence of a second ciliate mitochondrial genome has allowed us to perform a detailed re-analysis of the previously determined *P. aurelia* mtDNA sequence, and particularly a re-evaluation of the numerous unidentified ORFs reported in the latter case (Pritchard *et al.*, 1990b). Sequence comparisons have revealed 13 ORFs in *T. pyriformis* mtDNA that have identifiable homologs in *P. aurelia* mtDNA (Table 3). In the case of five additional Tetrahymena ORFs, there are ORFs of similar size at the same location in Paramecium mtDNA, relative to flanking genes (Table 4). This raises the possibility that these positionally equivalent ORFs are also homologous, albeit so highly diverged that this inference cannot be made convincingly from sequence comparisons. If we exclude these positional equivalents, only three (*P. aurelia*) or four (*T. pyriformis*) ORFs seem to be truly unique in the two mtDNAs. Because of the highly biased codon usage pattern in the Tetrahymena mtDNA, codon signatures provide a strong indication that both the ciliate-specific and unique ORF classes (with the possible exceptions of *ymf71* and *ymf73*; see Results) are functional genes in Tetrahymena.

A very similar suite of genes is present in *T. pyriformis* and *P. aurelia* mtDNAs. Identified protein-coding genes are a subset of those present in plant (e.g. *M. polymorpha*) and other protist (e.g. *A. castellanii*) mtDNAs. In view of the extreme sequence divergence displayed by identified protein genes, our working hypothesis is that at least some of the currently unidentified ciliate-specific ORFs may be highly diverged, and therefore unrecognizable, versions of ''missing'' respiratory chain and/or ribosomal protein genes. If this assumption turns out to be correct, then the apparent differences in ciliate mitochondrial gene content would further diminish compared to other protist mitochondrial genomes.

In contrast to protein-coding genes, few tRNA genes are encoded by both Tetrahymena and Paramecium mtDNAs (seven and four distinct *trn* genes, respectively). Although it is possible that additional, structurally peculiar tRNA genes remain to be identified in both genomes, we argue that this is unlikely, for several reasons. First, a high proportion of both genomes (96 %, *T. pyriformis*; 87 %, *P. aurelia*) already has an assigned coding function, and intergenic spacer sequences are relatively short, often much shorter than the expected length of a tRNA gene. Hence, little of the genome remains to accommodate additional unidentified genes, although the unsequenced end of *P. aurelia* mtDNA could conceivably harbor a few more tRNA genes (Pritchard *et al.*, 1990b). Second, except for a $tRNA^{Met}$ gene, identified *trn* genes predict tRNA secondary structures that deviate minimally from that of a conventional tRNA. Thus, mtDNA-encoded tRNA genes would be expected to be readily identifiable. Third, in the case of Tetrahymena, only a limited set of mtDNA-encoded tRNA species has been identified by direct probing of the genome with isolated mitochondrial tRNAs (Suyama, 1986), and this set accounts for all of the tRNA genes identified in *T. pyriformis* mtDNA. As in other mitochondrial systems where the mitochondrial genome encodes an insufficient number of tRNAs to support protein synthesis, the deficiency in mtDNA-encoded tRNAs must be made up by nucleus-encoded tRNAs imported from the cytosol.

Comparison of the complete mtDNA sequences of two ciliates has highlighted novel aspects of mitochondrial genome and gene structure that deserve further investigation from the perspective of gene expression. The split *rns* gene and split and rearranged *rnl* gene in *T. pyriformis* mtDNA have previously been investigated in detail (Schnare *et al.*, 1986; Heinonen *et al.*, 1987, 1990). Comparison of the *P. aurelia* and *T. pyriformis* mitochondrial genome sequences has revealed that the *nad1* gene is also split and rearranged, with N-terminal and C-terminal portions transcribed from different DNA strands. In addition, we have found that in one or both ciliate mitochondrial genomes, a number of protein-coding genes lack AUG initiation codons, raising questions about how these genes are expressed. These and other unique aspects of gene expression in ciliate mitochondria are explored in depth in the accompanying paper (Edqvist *et al.*, 2000).

In phylogenetic trees based on a concatenated dataset of mtDNA-encoded protein sequences, a monophyletic origin of the mitochondrial genome is strongly supported (as it is also in trees based on mitochondrial SSU rRNA sequences; Gray & Spencer, 1996). This result is notable in view of the fact that the ciliate sequences are on extraordinarily long branches (Figure 3), a manifestation of their exceptionally high rate of sequence divergence. The similarities in gene content between ciliate and other mtDNAs additionally support a single mitochondrial origin. Peculiarities in ciliate mitochondrial genome and gene structure are best rationalized by assuming a high degree of evolutionary divergence within the ciliate mitochondrial lineage, rather than by invoking a separate origin of the ciliate mitochondrial genome. Although our phylogenetic analyses suggest a specific evolutionary link between ciliate and apicomplexan (Plasmodium) mitochondrial genomes, and between these and the mtDNAs of certain amoeboid protozoa (Figure 3), these links have to be regarded as tentative due to the relatively long branches in this part of the tree, absence of data (e.g. *cox2* and *cox3* have not been identified in apicomplexan and ciliate mtDNAs, respectively), and relatively poor taxon sampling for ciliates, slime molds and rhizopod amoebae.

Materials and Methods

Culture conditions and preparation of nucleic acids from *Tetrahymena pyriformis*

Tetrahymena pyriformis, amicronucleate strain ST (kindly provided by Y. Suyama, Department of Biology, University of Pennsylvania) was grown at 28 °C with constant shaking in 500 ml of Neff's medium (Leick & Plesner, 1968). The culture was chilled on ice for 10-15 minutes, following which cells were collected by centrifugation at 2000 rpm for five minutes (IEC CR-6000 centrifuge, #219 rotor), resuspended, and centrifuged at 2000 rpm for five minutes through a layer of ice-cold homogenizing medium (0.35 M sucrose, 10 mM Tris-HCl (pH 7.2), 2 mM EDTA) (Schnare *et al*., 1986). Cells resuspended in 50 ml of homogenizing medium were disrupted by three passages through a hand emulsifier with the nozzle unscrewed 2.5 turns (Cowan & Young, 1978). Unbroken cells and debris were removed by centrifugation at 3000 rpm for five minutes (this and subsequent centrifugations were carried out in an IEC B-20A centrifuge using a #870 rotor). Mitochondria were recovered from the resulting supernatant by centrifugation at 8000 rpm for five minutes and washed twice by resuspension in 50 ml of homogenizing medium followed by centrifugation (five minutes, 8000 rpm). The final mitochondrial pellet was resuspended in 10 ml of 0.15 M NaCl, 0.1 M EDTA (pH 9.0) (Upholt & Borst, 1974) at room temperature. Ten ml of 4 % (w/v) SDS in the same buffer were added and the mitochondrial lysate was then extracted three times with phenol/cresol (Kirby, 1965) at room temperature. Nucleic acids were precipitated by addition of two volumes of 95 % (v/v) ethanol.

Cloning of mtDNA

Mitochondrial DNA was physically fragmented by nebulization (Okpodu *et al*., 1994). After fractionation by electrophoresis in an agarose gel, the random, size-selected fragments of mtDNA (500-1000 bp and 1000-3000 bp) were incubated with *Escherichia coli* DNA polymerase I (Klenow fragment) and phage T7 DNA polymerase to generate blunt ends and then cloned into the *Sma*I site of pBluescriptII KS$^+$ (Stratagene). Recombinant plasmids containing mtDNA inserts were identified by colony hybridization using total mtDNA as a probe. Clones contained in this random library encompassed the entire *T. pyriformis* mitochondrial genome except for three regions of 150, 1000 and 1500 bp. Three pairs of oligonucleotide primers were synthesized to permit amplification of these non-cloned regions by polymerase chain reaction using *Taq* polymerase. The resulting PCR products were cloned as above and sequenced.

Sequencing strategies

DNA sequencing was performed by the dideoxy chain termination method (Sanger *et al*., 1977). Single-stranded DNA was obtained by superinfection of recombinant clones with helper phage K07 (Vieira & Messing, 1987). To permit extended reading of sequences, high-resolution polyacrylamide gel electrophoresis was performed (Lang & Burger, 1990). Both strands were sequenced, and in regions where clone coverage was inadequate, specific sequencing primers (five in total) were synthesized for use in primer walking.

Reverse transcriptase sequencing of *T. pyriformis* mitochondrial rRNA was carried out as described (Lonergan & Gray, 1993), using oligonucleotide primers complementary to different regions of the mature rRNA sequence.

Analysis of sequence data

Sequences were read manually and assembled using the XBAP package (Dear & Staden, 1991). Sequence analysis was performed on DOS-based 486 microcomputers or SUN workstations using software developed by two of the authors (Lang & Burger, 1986), as well as with tools included in the Staden (1996) and MicroGenie (Queen & Korn, 1984) sequence analysis packages. The FASTA program (Pearson, 1990) was used for searches in local databases; sequence similarity searches were also performed at the National Center for Biotechnology Information (NCBI), using the BLAST network service (Altschul *et al*., 1990). Sequence similarities were evaluated (e.g. in defining the class of ciliate-specific ORFs; see below) by employing the RDF2 program (Pearson, 1990). Parameters employed in the latter analysis were ktup = 2, 1000 random uniform shuffles. The CLUSTAL V (Higgins & Sharp, 1989) and PIMA (Smith & Smith, 1992) programs were used for multiple protein alignments. Both programs were managed in the GDE (Genetic Data Environment) package (Smith *et al*., 1994). BLAST searches were conducted with the utility BBLAST (Batch Blast Search Tool; bionet.software, Message-ID: D2KnEF.4D3@cc.umontreal.ca) and large-scale output was screened with TBOB (Text-based Blast Output Browser; bionet.software, Message-ID: CpnMwI.LAH@cc.umontreal.ca). A number of other programs, including multiple sequence file manipulation, pre-processing and conversion utilities for XBAP, FASTA and GDE have been developed in the Sequencing Unit of the Organelle

Genome Megasequencing Program (OGMP). These utilities as well as BBLAST and TBOB are available through the OGMP website (URL http://megasun.bch.umontreal.ca/ogmpid.html).

The complete sequence of *T. pyriformis* mtDNA is deposited in GenBank (accession number AF160864). Sequences encompassing the rRNA genes and flanking regions have been determined independently in the laboratory of M.W.G. and have been published (Schnare *et al*., 1986; Heinonen *et al*., 1987, 1990). These sequences (GenBank accession numbers M12714, M58010, M58011) were completely re-determined here. In light of the *T. pyriformis* mtDNA sequence presented here, we also re-analyzed the published sequence of *P. aurelia* mtDNA (Pritchard *et al*., 1990b; GenBank acc. no. X15917) and have corrected and updated the accompanying annotation. This revised file as well as the one for the *T. pyriformis* mtDNA sequence are available through the Organelle Genome Database Project (GOBASE; Korab-Laskowski *et al*., 1997; URL http://megasun.bch.umontreal.ca/gobase).

Phylogenetic analysis

Phylogenetic tree construction employed the most recent versions of the programs PROTDIST, FITCH, NEIGHBOUR, PROTML and PUZZLE (Felsenstein, 1993; Fitch & Margoliash, 1967; Saitou & Nei, 1987; Strimmer & von Haeseler, 1996). Bootstrap and likelihood estimations were performed according to Felsenstein (1985) and Kishino *et al*. (1990), respectively.

Acknowledgments

We thank Y. Suyama for kindly providing the *T. pyriformis* strain used in this study, T. Y. K. Heinonen for advice on isolation of mitochondria, and members of the Gray laboratory for critical comment. This work was supported by grants SP-34 from the Medical Research Council of Canada and GO-12323 from the Canadian Genome Analysis and Technology Program. The study was also greatly assisted by a generous donation of computer equipment from Sun Microsystems. G.B. is an Associate and B.F.L. and M.W.G. are Fellows in the Program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR). Salary support from CIAR (M.W.G., B.F.L., G.B.) and the Walter C. Sumner Foundation (S.J.G.) is gratefully acknowledged.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Arnberg, A. C., van Bruggen, E. F. J., Borst, P., Clegg, R. A., Schutgens, R. B. H., Weijers, P. J. & Goldbach, R. W. (1975). Mitochondrial DNA of *Tetrahymena pyriformis* strain ST contains a long terminal duplication-inversion. *Biochim. Biophys. Acta,* **383**, 359-369.

Baldauf, S. L. & Doolittle, W. F. (1997). Origin and evolution of the slime molds. *Proc. Natl Acad. Sci. USA,* **28**, 12007-12012.

Budin, K. & Philippe, H. (1998). New insights into the phylogeny of eukaryotes based on ciliate Hsp70 sequences. *Mol. Biol. Evol.* **15**, 943-956.

Burger, G., Plante, I., Lonergan, K. M. & Gray, M. W. (1995). The mitochondrial DNA of the amoeboid protozoon, *Acanthamoeba castellanii*: complete sequence, gene content and genome organization. *J. Mol. Biol.* **245**, 522-537.

Cavalier-Smith, T. (1993). Kingdom protozoa and its 18 phyla. *Microbiol. Rev.* **57**, 953-994.

Chiu, N., Chiu, A. & Suyama, Y. (1975). Native and imported transfer RNA in mitochondria. *J. Mol. Biol.* **99**, 37-50.

Commission on Plant Gene Nomenclature (1993). A nomenclature for sequenced plant genes. *Plant Mol. Biol. Rep.* **11**, 291-316.

Cowan, A. E. & Young, P. G. (1978). The formation of several mitochondrial enzymes during the cell cycle in heat-synchronized *Tetrahymena pyriformis* ST. *Exp. Cell Res.* **112**, 79-87.

Cummings, D. J. (1992). Mitochondrial genomes of the ciliates. *Int. Rev. Cytol.* **141**, 1-64.

Dear, S. & Staden, R. (1991). A sequence assembly and editing program for efficient management of large projects. *Nucl. Acids Res.* **19**, 3907-3911.

Edqvist, J., Burger, G. & Gray, M. W. (2000). Expression of mitochondrial protein-coding genes in *Tetrahymena pyriformis*. *J. Mol. Biol.* **296**, 367-379.

Feagin, J. E. (1994). The extrachromosomal DNAs of apicomplexan parasites. *Annu. Rev. Microbiol.* **48**, 81-104.

Feagin, J. E., Werner, E., Gardner, M. J., Williamson, D. H. & Wilson, R. J. M. (1992). Homologies between the contiguous and fragmented rRNAs of the two *Plasmodium falciparum* extrachromosomal DNAs are limited to core sequences. *Nucl. Acids Res.* **20**, 879-887.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution,* **39**, 783-791.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* **22**, 521-565.

Felsenstein, J. (1993). *Phylip (Phylogeny Inference Package) Versions 3.5c and 3.6*, Department of Genetics, University of Washington, Seattle.

Fitch, W. M. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science,* **155**, 279-284.

Flavell, R. A. & Jones, I. G. (1970). Mitochondrial deoxyribonucleic acid from *Tetrahymena pyriformis* and its kinetic complexity. *Biochem. J.* **116**, 811-817.

Gajadhar, A. A., Marquardt, W. C., Hall, R., Gunderson, J., Ariztia-Carmona, E. V. & Sogin, M. L. (1991). Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. *Mol. Biochem. Parasitol.* **45**, 147-154.

Goldbach, R. W., Arnberg, A. C., van Bruggen, E. F. J., Defize, J. & Borst, P. (1977). The structure of *Tetrahymena pyriformis* mitochondrial DNA. I. Strain differences and occurrence of inverted repetitions. *Biochim. Biophys. Acta,* **477**, 37-50.

Goldbach, R. W., Borst, P., Bollen-de Boer, J. E. & Van Bruggen, E. F. J. (1978a). The organization of ribosomal RNA genes in the mitochondrial DNA of *Tetrahymena pyriformis* strain ST. *Biochim. Biophys. Acta,* **521**, 169-186.

Goldbach, R. W., Bollen-de Boer, J. E., Van Bruggen, E. F. J. & Borst, P. (1978b). Conservation of the sequence and position of the ribosomal RNA genes

in *Tetrahymena pyriformis* mitochondrial DNA. *Biochim. Biophys. Acta,* **521**, 187-197.

Gray, M. W. (1992). The endosymbiont hypothesis revisited. *Int. Rev. Cytol.* **141**, 233-357.

Gray, M. W. & Spencer, D. F. (1996). Organellar evolution. In *Evolution of Microbial Life* (Roberts, D. M., Sharp, P., Alderson, G. & Collins, M., eds), pp. 109-126, Cambridge University Press, UK.

Gray, M. W., Lang, B. F., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T. G., Plante, I., Rioux, P., Saint-Louis, D., Zhu, Y. & Burger, G. (1998). Genome structure and gene content in protist mitochondrial DNAs. *Nucl. Acids Res.* **26**, 865-878.

Heinonen, T. Y. K., Schnare, M. N., Young, P. G. & Gray, M. W. (1987). Rearranged coding segments, separated by a transfer RNA gene, specify the two parts of a discontinuous large subunit ribosomal RNA in *Tetrahymena pyriformis*. *J. Biol. Chem.* **262**, 2879-2887.

Heinonen, T. Y. K., Schnare, M. N. & Gray, M. W. (1990). Sequence heterogeneity in the duplicate large subunit ribosomal RNA genes of *Tetrahymena pyriformis* mitochondrial DNA. *J. Biol. Chem.* **265**, 22336-22341.

Hekele, A. & Beier, H. (1991). Nucleotide sequence and functional characterization of a mitochondrial $tRNA^{Trp}$ from *Tetrahymena thermophila*. *Nucl. Acids Res.* **19**, 1941.

Higgins, D. G. & Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *Comput. Appl. Biosci.* **5**, 151-153.

Inagaki, Y., Hayashi-Ishimaru, Y., Ehara, M., Igarashi, I. & Ohama, T. (1997). Algae or protozoa: phylogenetic position of euglenophytes and dinoflagellates as inferred from mitochondrial sequences. *J. Mol. Evol.* **45**, 295-300.

Kairo, A., Fairlamb, A. H., Gobright, E. & Nene, V. (1994). A 7.1 kb linear DNA molecule of *Theileria parva* has scrambled rDNA sequences and open reading frames for mitochondrially encoded proteins. *EMBO J.* **13**, 898-905.

Kishino, H., Miyata, T. & Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31**, 151-160.

Kirby, K. S. (1965). Isolation and characterization of ribosomal ribonucleic acids. *Biochem. J.* **96**, 266-269.

Korab-Laskowska, M., Rioux, P., Brossard, N., Littlejohn, T. G., Gray, M. W., Lang, B. F. & Burger, G. (1998). The organelle genome database project (GOBASE). *Nucl. Acids Res.* **26**, 138-144.

Labriola, J., Weiss, I., Zapatero, J. & Suyama, Y. (1987). Unexpectedly long 14 S ribosomal RNA gene in *Tetrahymena* mitochondria. *Curr. Genet.* **11**, 529-536.

Lang, B. F. & Burger, G. (1986). A collection of programs for nucleic acid and protein analysis, written in FORTRAN 77 for IBM-PC compatible microcomputers. *Nucl. Acids Res.* **14**, 455-465.

Lang, B. F. & Burger, G. (1990). A rapid, high resolution DNA sequencing gel system. *Anal. Biochem.* **188**, 176-180.

Lang, B. F., Burger, G., O'Kelly, C. J., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M. & Gray, M. W. (1997). An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature,* **387**, 493-497.

Leick, V. & Plesner, P. (1968). Formation of ribosomes in *Tetrahymena pyriformis*. *Biochim. Biophys. Acta,* **169**, 398-408.

Lonergan, K. M. & Gray, M. W. (1993). Editing of transfer RNAs in *Acanthamoeba castellanii* mitochondria. *Science,* **259**, 812-816.

McIntosh, M. T., Srivastava, R. & Vaidya, A. B. (1998). Divergent evolutionary constraints on mitochondrial and nuclear genomes of malaria parasites. *Mol. Biochem. Parasitol.* **95**, 69-80.

Middleton, P. G. & Jones, I. G. (1987). The terminus of *Tetrahymena pyriformis* mtDNA contains a tandemly repeated 31 bp sequence. *Nucl. Acids Res.* **15**, 855.

Morin, G. B. & Cech, T. R. (1988a). Mitochondrial telomeres: surprising diversity of repeated telomeric DNA sequences among six species of *Tetrahymena*. *Cell,* **52**, 367-374.

Morin, G. B. & Cech, T. R. (1988b). Phylogenetic relationships and altered genome structures among *Tetrahymena* mitochondrial DNAs. *Nucl. Acids Res.* **16**, 327-346.

Norman, J. E. & Gray, M. W. (1997). The cytochrome oxidase subunit 1 gene (*cox1*) from the dinoflagellate, *Crypthecodinium cohnii*. *FEBS Letters,* **413**, 333-338.

Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Takemura, M., Nozato, N., Akashi, K., Kanegae, T., Ogura, Y., Kohchi, T. & Ohyama, K. (1992). Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA. A primitive form of plant mitochondrial genome. *J. Mol. Biol.* **223**, 1-7.

Okpodu, C. M., Robertson, D., Boss, W. F., Togasaki, R. K. & Surzycki, S. J. (1994). Rapid isolation of nuclei from carrot suspension culture cells using a BioNebulizer. *Biotechniques,* **16**, 154-159.

Orr, A. T., Rabets, J. C., Horton, T. L. & Landweber, L. F. (1997). RNA editing missing in mitochondria. *RNA,* **3**, 335-336.

Paquin, B., Laforest, M.-J., Forget, L., Roewer, I., Wang, Z., Longcore, J. & Lang, B. F. (1997). The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. *Curr. Genet.* **31**, 380-395.

Patterson, D. J. & Sogin, M. L. (1992). Eukaryote origins and protistan diversity. In *The Origin and Evolution of the Cell* (Hartman, H. & Matsuno, K., eds), pp. 13-46, World Scientific Publishing Co. Pte. Ltd., Singapore.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63-98.

Pritchard, A. E., Sable, C. L., Venuti, S. E. & Cummings, D. J. (1990a). Analysis of NADH dehydrogenase proteins, ATPase subunit 9, cytochrome *b*, and ribosomal protein L14 encoded in the mitochondrial DNA of *Paramecium*. *Nucl. Acids Res.* **18**, 163-171.

Pritchard, A. E., Seilhamer, J. J., Mahalingam, R., Sable, C. L., Venuti, S. E. & Cummings, D. J. (1990b). Nucleotide sequence of the mitochondrial genome of *Paramecium*. *Nucl. Acids Res.* **18**, 173-180.

Queen, C. & Korn, L. J. (1984). A comprehensive sequence analysis program for the IBM personal computer. *Nucl. Acids Res.* **12**, 581-599.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA,* **74**, 5463-5467.

Schlegel, M. (1991). Protist evolution and phylogeny as discerned from small subunit ribosomal RNA sequence comparisons. *Eur. J. Protistol.* **27**, 207-219.

Schnare, M. N., Heinonen, T. Y. K., Young, P. G. & Gray, M. W. (1985). Phenylalanine and tyrosine transfer RNAs encoded by *Tetrahymena pyriformis* mitochondrial DNA: primary sequence, post-transcriptional modifications, and gene localization. *Curr. Genet.* **9**, 389-393.

Schnare, M. N., Heinonen, T. Y. K., Young, P. G. & Gray, M. W. (1986). A discontinuous small subunit ribosomal RNA in *Tetrahymena pyriformis* mitochondria. *J. Biol. Chem.* **261**, 5187-5193.

Schnare, M. N., Greenwood, S. J. & Gray, M. W. (1995). Primary sequence and post-transcriptional modification pattern of an unusual mitochondrial $tRNA^{Met}$ from *Tetrahymena pyriformis*. *FEBS Letters,* **362**, 24-28.

Sharma, I., Pasha, S. T. & Sharma, Y. D. (1998). Complete nucleotide sequence of the *Plasmodium vivax* 6 kb element. *Mol. Biochem. Parasitol.* **97**, 259-263.

Smith, R. F. & Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng.* **5**, 35-41.

Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W. & Gillevet, P. M. (1994). The genetic data environment, an expandable GUI for multiple sequence analysis. *Comput. Appl. Biosci.* **10**, 671-675.

Sogin, M. L. (1989). Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. *Am. Zool.* **29**, 487-499.

Sogin, M. L. (1991). Early evolution and the origin of eukaryotes. *Curr. Opin. Genet. Dev.* **1**, 457-463.

Sogin, M. L. (1997). History assignment: when was the mitochondrion founded? *Curr. Opin. Genet. Dev.* **7**, 792-799.

Staden, R. (1996). The Staden sequence analysis package. *Mol. Biotechnol.* **5**, 233-241.

Steinberg, S. & Cedergren, R. (1994). Structural compensation in atypical mitochondrial tRNAs. *Nature Struct. Biol.* **1**, 507-510.

Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964-969.

Suyama, Y. (1982). Native and imported tRNAs in *Tetrahymena* mitochondria: evidence for their involvement in intramitochondrial translation. In *Mitochondrial Genes* (Slonimski, P., Borst, P. & Attardi, G., eds), pp. 449-455, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Suyama, Y. (1985). Nucleotide sequences of three tRNA genes encoded in *Tetrahymena* mitochondrial DNA. *Nucl. Acids Res.* **13**, 3273-3284.

Suyama, Y. (1986). Two dimensional polyacrylamide gel electrophoresis analysis of *Tetrahymena* mitochondrial tRNA. *Curr. Genet.* **10**, 411-420.

Suyama, Y. & Jenney, F. (1989). The $tRNA^{glu}$ (anticodon TTC) gene and its upstream sequence coding for a homolog of the *E. coli* large ribosome-subunit protein L14 in the *Tetrahymena* mitochondrial genome. *Nucl. Acids Res.* **17**, 803.

Suyama, Y. & Miura, K. (1968). Size and structural variations of mitochondrial DNA. *Proc. Natl Acad. Sci. USA,* **60**, 235-242.

Suyama, Y., Jenney, F. & Okawa, N. (1987). Two transfer RNA sequences abut the large ribosomal RNA gene in *Tetrahymena* mitochondrial DNA: $tRNA^{leu}$ (anticodon UAA) and $tRNA^{met}$ (anticodon CAU). *Curr. Genet.* **11**, 327-330.

Thöny-Meyer, L. (1997). Biogenesis of respiratory cytochromes in bacteria. *Microbiol. Mol. Biol. Rev.* **61**, 337-376.

Upholt, W. B. & Borst, P. (1974). Accumulation of replicative intermediates of mitochondrial DNA in *Tetrahymena pyriformis* grown in ethidium bromide. *J. Cell. Biol.* **61**, 383-397.

Vaidya, A. B., Akella, R. & Suplick, K. (1989). Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malarial parasite. *Mol. Biochem. Parasitol.* **35**, 97-108.

Vaidya, A. B., Lashgari, M. S., Pologe, L. G. & Morrisey, J. (1993). Structural features of *Plasmodium* cytochrome *b* that may underlie susceptibility to 8-aminoquinolines and hydroxynaphthoquinones. *Mol. Biochem. Parasitol.* **58**, 33-42.

Vieira, J. & Messing, J. (1987). Production of single-stranded plasmid DNA. *Methods Enzymol.* **153**, 3-11.

Wilson, R. J. & Williamson, D. H. (1997). Extrachromosomal DNA in the Apicomplexa. *Microbiol. Mol. Biol. Rev.* **61**, 1-16.

Ziaie, Z. & Suyama, Y. (1987). The cytochrome oxidase subunit I gene of *Tetrahymena*: a 57 amino acid NH_2-terminal extension and a 108 amino acid insert. *Curr. Genet.* **12**, 357-368.

Edited by M. Yaniv

(Received 11 October 1999; accepted 14 January 2000)