Nucleic Acids Research, 2003, Vol. 31, No. 6 ***1673–1682***
DOI: 10.1093/nar/gkg270

Complete sequence of the mitochondrial genome of *Tetrahymena thermophila* and comparative methods for identifying highly divergent genes

Clifford F. Brunk*, Louis C. Lee, Anne B. Tran and Jinliang Li[1]

Department of Organismic Biology, Ecology and Evolution, University of California–Los Angeles, Los Angeles, CA 90095-1606, USA and [1]Laragen Inc., 10755 Venice Boulevard, Los Angeles, CA 90034, USA

Received November 24, 2002; Revised and Accepted January 16, 2003 DDBJ/EMBL/GenBank accession no. AF396436

ABSTRACT

The complete sequence of the mitochondrial genome of *Tetrahymena thermophila* has been determined and compared with the mitochondrial genome of *Tetrahymena pyriformis*. The sequence similarity clearly indicates homology of the entire *T.thermophila* and *T.pyriformis* mitochondrial genomes. The *T.thermophila* genome is very compact, most of the intergenic regions are short (only three are longer than 63 bp) and comprise only 3.8% of the genome. The *nad9* gene is tandemly duplicated in *T.thermophila.* Long terminal inverted repeats and the *nad9* genes are undergoing concerted evolution. There are 55 putative genes: three ribosomal RNA genes, eight transfer RNA genes, 22 proteins with putatively assigned functions and 22 additional open reading frames of unknown function. In order to extend indications of homology beyond amino acid sequence similarity we have examined a number of physico-chemical properties of the mitochondrial proteins, including theoretical pI, molecular weight and particularly the predicted transmembrane spanning regions. This approach has allowed us to identify homologs to *ymf58* (*nad4L*), *ymf62* (*nad6*) and *ymf60* (*rpl6*).

INTRODUCTION

The mitochondrial DNA of *Tetrahymena pyriformis* was among the first protist mitochondrial DNAs characterized (1,2). The *Tetrahymena* mitochondrial DNA is linear (~47 kb in length). An examination of the mitochondrial genome of *Tetrahymena* reveals several unusual properties. Only seven different tRNAs are coded by the mitochondria. The remaining mitochondrial tRNAs are of nuclear origin and are imported into the mitochondria (3–7). The terminal ends are long inverted repeats that contain the large subunit ribosomal RNA (*LSU rRNA*) genes and the *leucine tRNA* genes (8,9). In the *Tetrahymena* mitochondrial DNA, both the *LSU rRNA* gene and the small subunit ribosomal RNA (*SSU rRNA*) genes are split. The *nad1* gene is also split into two transcriptional units (7,10–12).

Most mitochondrial genomes contain genes that code for proteins involved in pathways of oxidative phosphorlyation and protein synthesis (13). In *T.pyriformis* there are 22 proteins with assigned function. Of these, 13 are involved in oxidative phosphorylation pathways, eight are ribosomal proteins and one is involved in cytochrome c biosynthesis. An additional 22 putative proteins are of unknown function (7).

Among ciliates, *Tetrahymena thermophila* and *T.pyriformis* are close evolutionary relatives, but within the more than 30 species comprising the *T.pyriformis* complex, *T.thermophila* and *T.pyriformis* are fairly distant relatives (14,15). Among all known organisms, *Tetrahymena malaccensis* is the most closely related species to *T.thermophila*. A detailed comparison of the mitochondrial genomes of *T.thermophila* and *T.pyriformis* allows the unequivocal identification of protein coding regions. The degree of divergence among these regions is indicative of the selective pressure acting on these various genes. In cases where the *T.thermophila* and *T.pyriformis* genome organization differs, we have examined the *T.malaccensis* mitochondrial sequences to gain further insight into the evolutionary histories of these mitochondrial genes.

Many ciliates, including *T.thermophila* and *T.pyriformis*, use a mitochondrial genetic code in which TAA is the sole termination codon, TGA codes for tryptophan and TAG is not used (13). The genes in the mitochondrial genome of *T.pyriformis* apparently use four initiation codons (ATA, ATT, TTG and GTG) in addition to the standard ATG (16). Polygenic transcriptional units are also indicated, but there is no evidence of RNA editing (16).

During the evolution of modern eukaryotes the mitochondrial genome has been dramatically modified in various different lineages. The vast majority of the genes present in the original endosymbiont have been lost from the mitochondrial genome. Most of these genes are simply gone from the cell, but a number of genes originally in the mitochondria have been transferred to the nucleus (17,18). In a few eukaryotic lineages, the mitochondria and the mitochondrial genome have been lost entirely (19,20). This process of elimination and transfer of mitochondrial genes has followed different evolutionary courses in different lineages.

*To whom correspondence should be addressed. Tel: +1 310 825 3114; Fax: +1 310 206 3987; Email: cbrunk@ucla.edu

The genes present in mitochondrial genomes fall into distinct categories: (i) rRNAs; (ii) tRNAs; (iii) ribosomal proteins; (iv) proteins involved in oxidative phosphorylation; (v) other proteins (not ribosomal proteins and not directly involved in oxidative phosphorylation); and (vi) open reading frames (ORFs) coding for putative proteins for which homologs of known function have not be identified (21). Among the putative mitochondrial protein coding genes, ~50% are involved in oxidative phosphorylation, 30% are ribosomal proteins, 7% are other proteins and 13% are of unknown function. Thus, there are a limited set of genes that are expected among the mitochondrial ORFs, which makes their identification considerably easier.

The ciliates have among the most rapidly evolving mitochondrial genomes as evidenced by the large number of putative proteins for which function cannot be readily assigned. By comparison, in another protist, *Reclinomonas americana*, 92% of the mitochondrial genes can be assigned functions (22).

The most common means of assigning putative function to mitochondrial genes is to identify proteins of known function that have highly similar amino acid sequences (23). It is generally assumed that homologous proteins will have homologous functions. Our objective is to assign tentative functions to as many of the putative protein coding genes in the *Tetrahymena* mitochondrial genome as possible. We have examined a series of characteristics, in addition to sequence similarity, that may also indicate homology between proteins. These include physico-chemical properties such as theoretical pI, molecular weight and particularly the predicted transmembrane (TM) spanning regions.

MATERIALS AND METHODS

A B strain of *T.thermophila* (SB210, provided by E. Orias, University of California, Santa Barbara), *T.malaccensis* (MP75, obtained from ATCC) and *T.pyriformis* (GL, obtained from H. Buhse, University of Illinois, Chicago) were grown in PYG medium (24,25). Mitochondria were isolated from *T.thermophila* and total DNA was prepared according to Brunk and Hanawalt (2). The mitochondrial DNA was sheared to fragments of ~3 kb in length using a shear point device (26). The ends of these fragments were filled in with T4 DNA polymerase and phosphorylated with polynucleotide kinase (27). Oligonucleotide adapters were ligated to the blunt ends of the mitochondrial DNA fragments and these fragments were directly cloned into pAMP 10 vectors (Life Technology, San Diego, CA). The sequences (~500 bp) at each end of the cloned fragments were determined. These sequences (~150 kb) were aligned using Phred, Phrap and Consed to form contigs (28–30). Clones that bridge the gaps between contigs were completely sequenced using specific primers to connect the contigs. The long inverted repeats in the terminal regions were independently PCR amplified using specific primers to achieve accurate consensus sequences for these regions. The final sequence has a >3-fold coverage of all regions and an error rate of 0.75 bases per 10 000 bases as estimated by the Consed program. The organization of the *nad9* genes (tandemly repeated) and the *atp9* gene in *T.malaccensis* were examined by cloning and sequencing these regions. The sequence of the *T.thermophila* mitochondrial genome was aligned with the sequence of the *T.pyriformis* mitochondrial genome, matching sequences of the ORFs at both the amino acid and nucleotide levels using Clustal X (31).

Table 1 shows the similarity metric (*z* score) used to measure the degree of sequence similarity between ORFs of mitochondrial genes in *T.thermophila* and *T.pyriformis* (32–34). The ORFs were translated into amino acid sequences, aligned and a normalized alignment score (NAS) was computed. The amino acid sequences were then randomized, aligned and additional NASs were computed. This process was repeated 1000 times and the mean and standard deviation of the resulting distribution was computed. The *z* score for two ORFs is the difference between their NAS value and the mean for the random distribution divided by the standard deviation of this distribution. This distribution approximates a Gumbel extreme value distribution (34). A *z* score >6 is expected by chance less than three times in 1000 trials, which is consistent with our random distributions. Thus, *z* scores >6 strongly indicate homology (34,35). A similar metric for sequence similarity was used in comparing nucleotide sequences.

The proportions of silent substitutions (Ks) and substitutions that change an amino acid (Ka) were calculated for each homolog in the *T.thermophila* and *T.pyriformis* mitochondrial genome and the Ks/Ka ratios were determined. The nine nucleotide changes possible for each codon were characterized as either silent or amino acid changing. These changes were averaged per nucleotide position. The actual nucleotide changes occurring between the sequences were characterized as well. When more that one nucleotide change occurred in a codon, all potential paths were considered on an unweighted basis and the changes were prorated per nucleotide position. The actual nucleotide changes per codon were normalized by the potential changes. These observed normalized nucleotide changes were corrected for multiple hits producing the Ks and Ka values (36,37). The Ks values and Ks/Ka ratios are shown in Table 1.

Each of the 45 ORFs in the *T.thermophila* mitochondrial genome was used as a query sequence with the NCBI BLAST program to produce a list of BLAST hits (38). ORFs with a list of BLAST hits containing numerous versions of the same protein (consensus protein) that had a low probability of occurring by chance were tentatively assigned the function of the consensus protein. Consensus proteins involved in oxidative phosphorylation and electron transport were compared to the homologs found in *Bos taurus* (cow), directly or via a chain of homology. Similarly, for the ribosomal proteins, *T.thermophila* was compared to *Saccharomyces cerevisiae* (yeast).

The *T.thermophila* ORFs were translated into amino acid sequences and aligned with the presumptive homolog using Clustal X (31). The aligned sequences were examined for the presence of sizable extensions at the N- or C-terminus and unaligned terminal additions >10 amino acids were trimmed. The *z* scores and trimming details for each homolog are shown in Table 2. The *nad1* split gene was treated as a single protein. Protein properties, including the theoretical pI, ratio of negative to positive residues, aliphatic index and the grand average of hydropathicity (GRAVY) for all of the *T.thermophila* ORFs were calculated using ProtParam (39). TM of the *T.thermophila* ORFs were predicted using TMHMM version 2 (40). The protein properties and number

Table 1. Comparative data between *T.thermophila* and *T.pyriformis*

Gene	*z* score	Ks/Ka	Ks	Gene length[a]	*T.pyriformis*[b]	*T.thermophila*[b]	Start position[c]
atp9	19.83	3.39	0.49	76	·	·	·
cob	74.41	34.27	0.92	426	(GTG)V	·	·
cox1	106.58	29.60	0.73	688	·	·	·
cox2	88.81	9.79	0.77	604	·	·	·
nad1_a	58.52	15.63	0.81	284	·	·	·
nad1_b	20.11	15.63	0.81	59	·	·	·
nad10	42.81	47.59	1.02	162	·	(TTG)L	·
nad2	37.89	10.66	0.94	178	·	·	·
nad3	28.85	8.69	1.07	121	·	·	·
nad4	83.64	15.30	1.02	505	·	·	·
nad5	83.87	4.20	0.93	750	(TTG)L	(TTG)L	·
nad7	82.82	17.70	0.69	442	·	·	·
nad9_1	43.30	16.33	1.40	198	·	·	·
nad9_2	43.30	16.15	1.32	198	·	·	·
rpl14	33.49	7.05	0.76	119	·	·	·
rpl16	34.23	8.65	0.97	143	(TTG)L	(TTG)L	·
rpl2	53.44	5.64	0.70	262	·	·	·
rps12	39.03	18.77	0.62	133	(TTG)L	(TTG)L	·
rps13	55.53	9.61	0.89	276	(ATA)I	(ATT)I	·
rps14	29.77	15.12	1.00	101	·	·	·
rps19	26.52	4.96	0.63	98	·	·	·
rps3	30.13	3.89	0.75	151	·	·	·
yejR	22.50	2.30	1.29	518	(ATA)I	(ATT)I	·
ymf56	28.46	10.76	0.69	97	(ATT)I	(ATA)I	–3
ymf57	24.72	5.81	0.52	100	·	·	·
ymf58	28.66	10.30	0.81	116	·	·	·
ymf59	29.92	5.10	1.01	152	·	(TTG)L	·
ymf60	32.06	4.32	1.18	179	·	·	·
ymf61	37.02	5.22	1.10	238	·	(ATA)I	·
ymf62	42.69	9.28	0.99	255	·	·	·
ymf63	42.12	5.89	1.33	276	·	·	·
ymf64	43.90	3.65	0.98	330	·	·	·
ymf65	58.81	14.55	1.65	364	(ATA)I	(ATT)I	+5
ymf66	63.20	5.23	0.84	446	(GTG)V	(GTG)V	·
ymf67	38.21	2.14	0.84	453	·	·	·
ymf68	82.71	6.78	0.84	594	(ATA)I	(ATT)I	·
ymf69	10.29	2.41	0.80	68	(ATT)I	(GTG)V	+3
ymf70	26.70	9.63	0.60	81	·	·	·
ymf71	6.86	1.66	0.78	101	(ATT)I	(ATT)I	–1
ymf72	22.20	2.76	0.98	117	·	(ATA)I	·
ymf73	28.97	8.05	1.84	159	(ATA)I	(ATA)I	·
ymf74	18.41	1.64	0.64	157	(ATA)I	(ATT)I	+2
ymf75	30.75	2.79	0.78	190	·	·	·
ymf76	46.41	2.60	0.72	405	·	·	–2
ymf77	59.93	1.61	0.79	1344	(ATA)I	·	+4

[a] Length of *T.thermophila* genes in amino acids.
[b] Alternate initiation codons and their respective amino acid. Dots represent standard initiation codon, ATG.
[c] Start position of *T.thermophila* genes relative to *T.pyriformis*. Dots represent the same position.

of TM regions are shown in Table 2. The *T.thermophila* ORFs were also used as queries for PRODOM and pFAM protein domain databases (41).

RESULTS

An alignment of the *T.thermophila* and *T.pyriformis* mitochondrial genomes indicates that these genomes are virtually identical with respect to gene content and order (Fig. 1), except for the tandem duplication of *nad9* in *T.thermophila*. Thus, we have labeled the *T.thermophila* ORFs with the same registered *ymf* designations used for the *T.pyriformis* genes (7).

The z scores calculated for each of the ORFs in *T.thermophila* and *T.pyriformis* (Table 1) all substantially exceed our criteria ($z > 6$) as an indicator of homology between the sequences. The RNA coding genes also show a high degree of sequence similarity as well.

Strong selective pressure on an ORF will favor silent substitutions (Ks) over nucleotide substitutions that lead to a change in an amino acid (Ka). The values for Ks (Table 1) indicate that, while there has been substantial nucleotide substitution in the *T.thermophila* and *T.pyriformis* mitochondrial genomes, the silent substitutions have not saturated the potential sites. The Ks/Ka ratio varies dramatically from gene to gene, providing an indication of the selective pressure on each protein.

As noted for the *T.pyriformis* mitochondrial genome, the *T.thermophila* mitochondrial genome appears to have a limited number of transcriptional units (7). To a first

Table 2. Similarity scores and physico-chemical values for identifiable mitochondrial ORFs

Protein	*z* score[a]	Organisms compared	Accession no.	Trim	Theoretical pI	Ratio of –/+ residues	Aliphatic index	GRAVY	TM
Rpl2		*Tetrahymena thermophila*	NP_149370	no	10.76	0.143	81.53	–0.477	0
	16.69	*Saccharomyces cerevisiae*	NP_010864	90N, 23C	10.78	0.351	84.30	–0.577	0
Rpl14		*Tetrahymena thermophila*	NP_149405	no	10.69	0.143	93.19	–0.257	0
	15.05	*Thermatoga maritima*	NP_229290	no	10.00	0.609	96.56	–0.220	0
	15.23	*Saccharomyces cerevisiae*	NP_012751	no	9.87	0.417	93.91	–0.106	0
Rpl16		*Tetrahymena thermophila*	NP_149386	no	10.99	0.083	88.60	–0.499	0
	9.36	*Phytophthora infestans*	NP_037626	no	10.68	0.206	87.99	–0.433	0
	14.93	*Saccharomyces cerevisiae*	gil463270	39N, 51C	10.32	0.413	76.03	–0.619	0
Rps12		*Tetrahymena thermophila*	NP_149373	10C	11.90	0.135	90.75	–0.656	0
	16.57	*Saccharomyces cerevisiae*	gil2119089	26N	11.18	0.207	71.31	–0.622	0
Rps13		*Tetrahymena thermophila*	NP_149367	154C	10.18	0.397	72.79	–0.825	0
	10.10	*Pseudomonas aeruginosa*	NP_252931	no	10.94	0.400	85.17	–0.681	0
	13.42	*Saccharomyces cerevisiae*	S53897	23C	10.43	0.320	84.62	–0.315	0
Rps14		*Tetrahymena thermophila*	NP_149376	no	10.60	0.077	90.79	–0.574	0
	6.72	*Deinococcus radiodurans*	NP_295832	no	10.84	0.360	60.28	–0.973	0
	14.85	*Escherichia coli*	NP_289868	no	11.16	0.400	72.57	–0.795	0
	10.77	*Saccharomyces cerevisiae*	NP_015492	no	11.04	0.292	86.52	–0.497	0
Rps19		*Tetrahymena thermophila*	NP_149369	no	10.58	0.192	81.53	–0.610	0
	8.54	*Nephroselmis olivacea*	AAF03188	no	11.26	0.200	70.62	–0.484	0
	8.28	*Saccharomyces cerevisiae*	NP_014435	no	10.54	0.400	79.34	–0.441	0
Ymf60 (Rpl6)		*Tetrahymena thermophila*	NP_149377	no	10.29	0.312	87.09	–0.413	0
	7.88	*Thermatoga maritima*	gil7440704	no	9.72	0.666	97.77	–0.311	0
	13.68	*Saccharomyces cerevisiae*	NP_012017	15N	9.90	0.594	99.63	–0.282	0
					10.58[b]	**0.318**	**84.04**	**–0.51**	**0**
Atp9		*Tetrahymena thermophila*	NP_149381	no	4.94	1.333	117.79	1.190	2
	6.99	*Bos taurus*	P32876	46N	9.89	0.500	106.25	0.637	2
Cob		*Tetrahymena thermophila*	NP_149395	no	5.45	1.429	95.18	0.358	9
	7.55	*Bos taurus*	gil117843	no	7.75	0.944	120.71	0.680	9
Cox1		*Tetrahymena thermophila*	NP_149402	38N	9.90	0.429	95.31	0.304	12
	27.72	*Bos taurus*	NP_008097	no	6.06	1.471	102.06	0.685	12
Cox2		*Tetrahymena thermophila*	NP_149397	10N, 24C	9.70	0.585	89.39	–0.465	2
	9.96	*Bos taurus*	NP_008098	no	4.78	2.083	108.19	0.247	2
Nad1		*Tetrahymena thermophila*	NP_149403	19N	6.81	1.000	130.41	1.094	8
	23.56	*Bos taurus*	NP_008095	56C	7.84	0.933	123.99	0.798	8
Nad2		*Tetrahymena thermophila*	NP_149374	no	9.52	0.333	145.11	0.923	4
	7.39	*Bos taurus*	NP_008096	180N	9.96	0.313	114.67	0.785	8
Nad3		*Tetrahymena thermophila*	NP_149389	no	4.54	2.250	122.40	0.969	3
	10.71	*Bos taurus*	gil5834947	no	4.50	2.000	128.17	0.863	3
Nad4		*Tetrahymena thermophila*	NP_149407	29N	8.43	0.846	123.74	0.943	12
	7.70	*Bos taurus*	gil128741	no	9.42	0.571	130.48	0.826	12
Nad5		*Tetrahymena thermophila*	NP_149396	166N	8.89	0.756	121.93	0.734	18
	16.60	*Bos taurus*	gil5834950	no	9.25	0.666	113.07	0.632	14
Nad7		*Tetrahymena thermophila*	NP_149375	no	8.70	0.880	83.78	–0.341	0
	44.93	*Bos taurus*	gil128846	19N	5.95	1.163	81.86	–0.338	0
Nad9_1		*Tetrahymena thermophila*	NP_149392	no	6.62	1.000	97.98	–0.308	0
	8.05	*Phytophthora megasperma*	AAA32025	16C / no	9.25	0.731	107.29	–0.144	0
	27.40	*Bos taurus*	AAA30663	56N, 10C	6.25	1.061	85.08	–0.300	0
Nad9_2		*Tetrahymena thermophila*	NP_149392	no	6.62	1.000	97.98	–0.308	0
	8.05	*Phytophthora megasperma*	AAA32025	16C / no	9.25	0.731	107.29	–0.144	0
	27.40	*Bos taurus*	AAA30663	56N, 10C	6.25	1.061	85.08	–0.300	0
Nad10		*Tetrahymena thermophila*	NP_149372	no	9.13	0.667	87.30	–0.048	0
	32.64	*Bos taurus*	gil1171865	60N	9.90	0.500	88.98	0.018	0
Ymf58 (Nad4L)		*Tetrahymena thermophila*	NP_149391	18C	5.10	1.167	139.57	0.778	3
	5.47	*Nephroselmis olivacea*	gil6066176	no	6.53	1.000	158.00	1.338	3
	8.57	*Penaeus monodon*	NP_038297	no	6.88	1.000	131.82	1.057	3
	11.51	*Bos taurus*	NP_008103	no	5.27	3.000	131.33	1.259	3
Ymf62 (Nad6)		*Tetrahymena thermophila*	NP_149404	19C	4.53	1.818	130.31	0.738	5
	5.61	*Porphyra purpurea*	NP_049300	no / 27C	8.47	0.818	147.80	1.066	5
	6.01	*Bos taurus*	NP_008106	no	4.15	2.600	115.66	1.031	5
					7.47[b]	**1.08**	**113.31**	**0.49**	**4.8**
YejR		*Tetrahymena thermophila*	NP_149387	87N, 256C	9.36	0.400	141.68	0.760	15
	8.14	*Pseudomonas putida*	AAC63584	no	8.13	0.889	107.31	0.568	4
	17.05	*Triticum aestivum*	S38799	252N, 79C	9.62	0.593	92.43	0.159	6

[a] *z* scores >6 strongly suggest homology (35).

[b] Characteristic signatures can be used to distinguish ribosomal proteins from those involved in oxidative phoshorylation and electron transport as indicated by their mean values (bold).

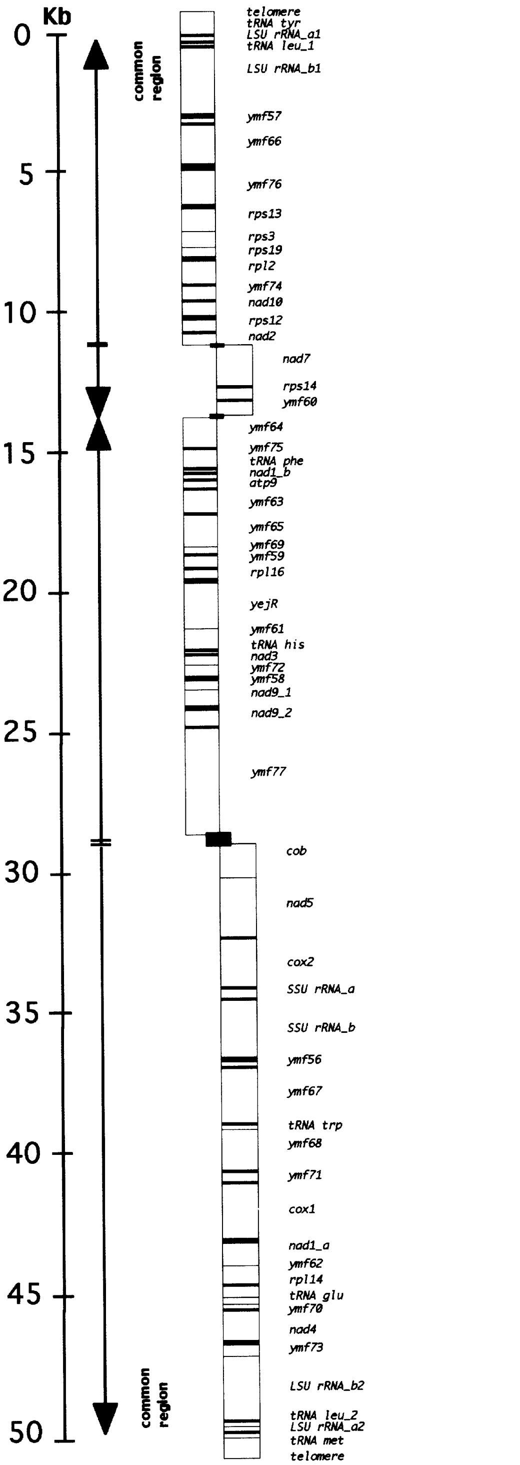


Figure 1. Gene map of the *T.thermophila* mitochondrial genome. Arrows denote direction of transcription. Genes are in white and intergenic regions are in black. The 5′ terminus is at the top.

Table 3. (**a**) Synonymous/non-synonymous nucleotide substitutions between *nad9* genes in *T.malaccencis*, *T.thermophila* and *T.pyriformis*[a], and (**b**) number of nucleotide substitutions between *T.thermophila* and *T.pyriformis* in the terminal repeat regions[b]

(**a**)	*T.mal*_1	*T.mal*_2	*T.the*_1	*T.the*_2	*T.pyr*
*T.mal*_1					
*T.mal*_2	3/2				
*T.the*_1	57/24	56/22			
*T.the*_2	56/23	55/21	9/1		
T.pyr	67/35	65/33	76/27	77/28	
(**b**)	*T.pyr* 5′	*T.pyr* 3′	*T.the* 5′	*T.the* 3′	
T.pyr 5′					
T.pyr 3′	4				
T.the 5′	230	228			
T.the 3′	231	231	6		

[a]The *nad9* genes in *T.malaccencis, T.thermophila* and *T.pyriformis* each have 94 nt.

[b]The terminal repeat regions in *T.thermophila* and *T.pyriformis* are 2684 and 2687 nt long, respectively.

approximation there appear to be two transcriptional units, transcribed from a nearly central bi-directional promoter set (Fig. 1). The largest intergenic gap in the genome (424 bp) is at the position of the putative central set of bi-directional promoters. A short transcriptional unit, including the three genes transcribed in the 5′→3′ direction, interrupts the large 3′→5′ transcriptional unit.

The most apparent difference between the two genomes is the tandem duplication of the *nad9* gene in the *T.thermophila* genome. *Tetrahymena malaccensis*, the closest known relative of *T.thermophila*, also has tandemly repeated *nad9* genes. The tandemly repeated genes are highly similar within each species, but diverge considerably between the two species (Table 3a). Similarily, the long inverted repeats in the terminal regions are virtually identical within a species, but diverge significantly between species (Table 3b).

The mitochondrial genes in *T.pyriformis* apparently use four initiation codons (ATA, ATT, TTG and GTG) in addition to the standard ATG (16). The vast majority of the *T.thermophila* mitochondrial putative proteins appear to start at the same relative position as their *T.pyriformis* homolog and use the same initiation codon. Seven putative proteins appear to initiate at somewhat different positions in *T.thermophila* and *T.pyriformis*. The most plausible initiation codons for *T.thermophila* were assigned based on three criteria: (i) their proximity to the 5′ end of the ORF; (ii) their juxtaposition to the assigned *T.pyriformis* initiation codons; and (iii) the utilization of only initiation codons assigned in *T.pyriformis.* In five cases, proteins that initiate with unconventional codons in one *Tetrahnymena* species are initiated in the other species at an identical nucleotide position with ATG (Table 1). This strongly supports the correct identification of these unconventional initiation codons in these sequences.

By far the largest gene in *T.thermophila* or *T.pyriformis* is *ymf77*. It is highly improbable that a region of >4000 nt would be devoid of TAA stop codons or TAG unused codons in both species by chance. A comparison of the nucleotide substitutions between *T.thermophila* and *T.pyriformis* for *ymf77* indicates that the selective pressure on this protein is probably not large (Ks/Ka = 1.61, Table 1).

Although most of the intergenic regions in *T.thermophila* and *T.pyriformis* are of similar length, the intergenic region following the *T.pyriformis atp9* gene is exceptionally large (95 bp) compared to the intergenic in *T.thermophila* (14 bp).

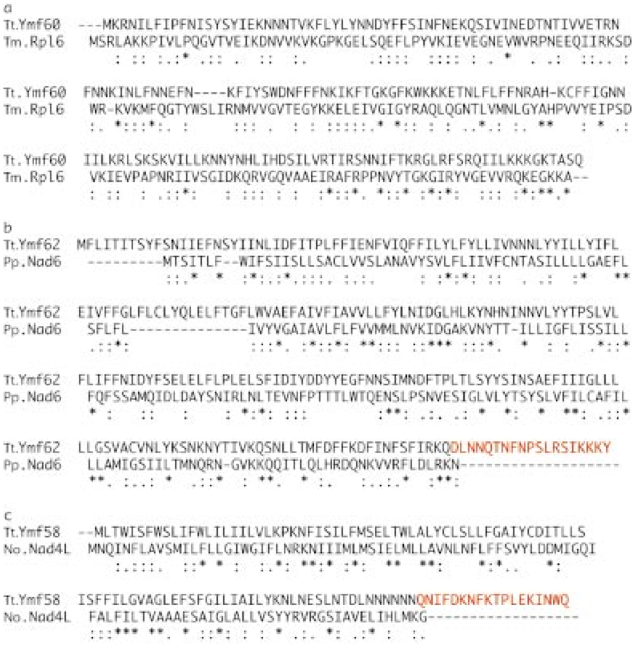
The *atp9* gene and its flanking regions from *T.malaccensis* were sequenced and compared with *T.thermophila* and *T.pyriformis*. This comparison indicates that the *T.malaccensis* and *T.thermophila atp9* genes and intergenic regions are virtually identical.

Our objective is to identify the probable function of each of the *Tetrahymena* genes based on their sequence similarity to proteins of known function. BLAST analysis provided initial identification of the *Tetrahymena* genes. Thirteen of the *T.thermophila* mitochondrial ORFs can be immediately assigned function by their similarity to standard proteins of *B.taurus* (Atp9, Cob, Cox1, Cox2, Nad1, Nad2, Nad3, Nad4, Nad5, Nad7 and Nad10) or *S.cerevisiae* (Rpl2 and Rps12) (Table 2). Two additional *T.thermophila* mitochondrial ORFs (Nad9_1 and Nad9_2) are linked via a chain of homology to *B.taurus* proteins and five further ORFs (Rpl14, Rpl16, Rps13, Rps14 and Rps19) are linked to *S.cerevisiae* proteins (Table 2). In this manner, the functions for 21 of the 45 *T.thermophila* ORFs can be reliably assigned by chains of sequence similarity comparison.

The remainder of the *T.thermophila* ORFs have BLAST hit lists that include a wide variety of proteins having different functions, with no clear consensus protein indicated. At this point the physico-chemical parameters for these protein sequences were examined for clues to their identity. The theoretical pI, ratio of negative to positive residues, aliphatic index and the GRAVY and TM were calculated for the ORFs with assigned function (Table 2).

TM are not present in the ribosomal proteins of *Tetrahymena* or in virtually any ribosomal proteins found in GenBank. Thus, the presence of TM strongly indicates that the ORF is not a ribosomal protein. Alternatively, most of the proteins involved in oxidative phosphorylation and electron transport have TM and many have multiple TM. However, some proteins of the oxidative phosphorylation NADH Fo complex (Nad5, Nad9 and Nad10) do not have TM. Generally, the theoretical pI for ribosomal proteins is high and rarely below 10, while the theoretical pI for proteins involved in oxidative phosphorylation and electron transport is low, usually below 8 and never above 10. The ratio of negative to positive residues for ribosomal proteins is low, rarely above 0.5, while this ratio is usually above 1.0 for proteins involved in oxidative phosphorylation and electron transport. Among the ribosomal proteins the aliphatic index is never above 100, while the aliphatic index for proteins involved in oxidative phosphorylation and electron transport is seldom below 100. The GRAVY for ribosomal proteins is always negative, averaging –0.51, while the GRAVY for proteins involved in oxidative phosphorylation and electron transport is seldom negative, averaging 0.49. Although these parameters are not totally independent of one another, taken together they present a valuable signature for ribosomal proteins and proteins involved in oxidative phosphorylation and electron transport.

The physico-chemical parameters for Ymf60 indicate that it is potentially a ribosomal protein (Table 2). It has no TM and all of the other parameters match those of ribosomes. Among the weak BLAST hits for Ymf60 were two hits for the large subunit ribosomal protein 6. We compared the sequence of Ymf60 to virtually all of the Rpl6 sequences in GenBank. The Rpl6 sequence from *Thermatoga maritima* was found to have sufficient similarity to Ymf60 to indicate homology (Fig. 2a), and a chain of homology linked Ymf60 to Rpl6 of *S.cerevisiae* (Table 2). Thus, Rpl6 function can be assigned to Ymf60 with confidence.



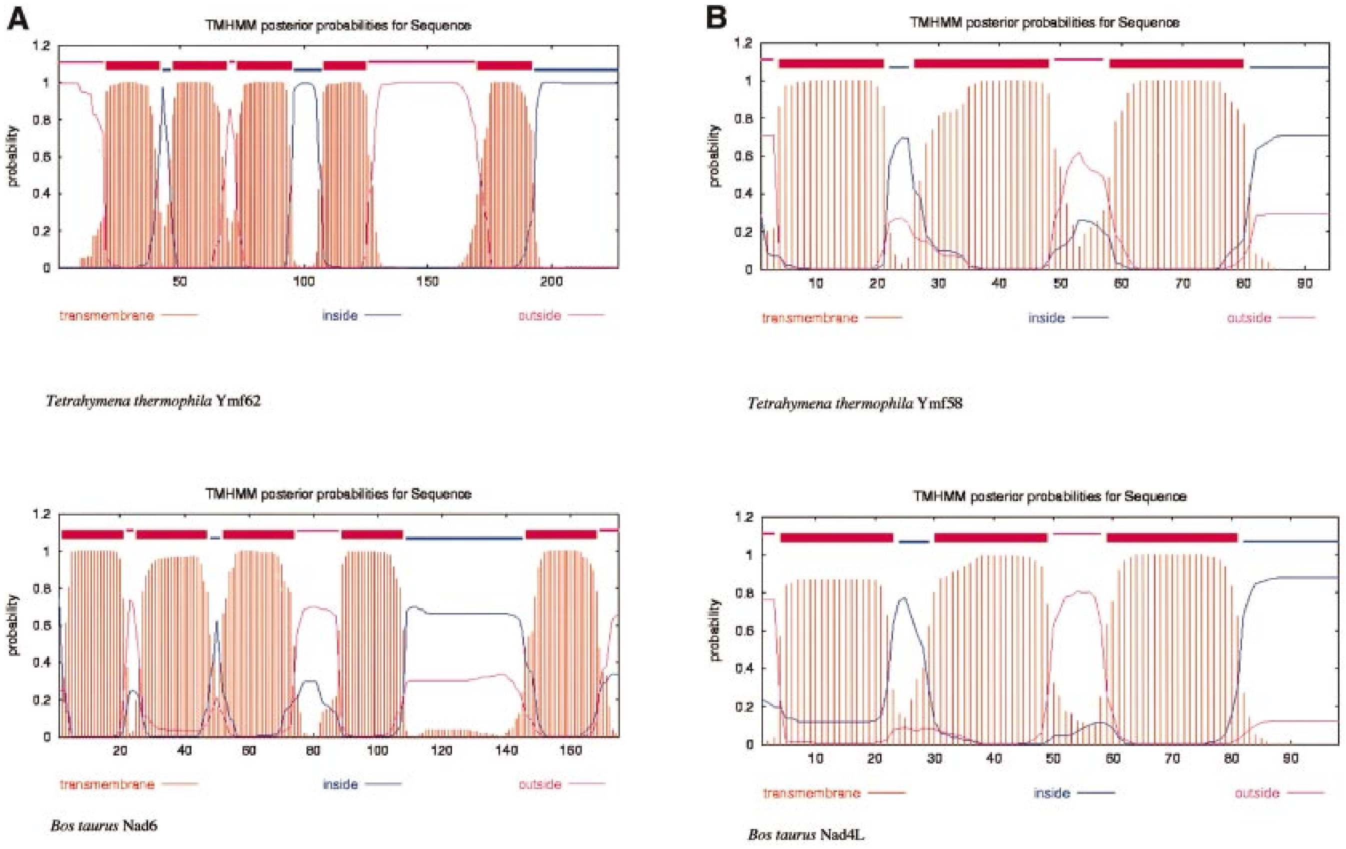
```
a
Tt.Ymf60   ---MKRNILFIPFNISYSYIEKNNNTVKFLYLYNNDYFFSINFNEKQSIVINEDTNTIVVETRN
Tm.Rpl6    MSRLAKKPIVLPQGVTVEIKDNVVKVKGPKGELSQEFLPYVKIEVEGNEVWVRPNEEQIIRKSD
              : :: :.:* .:: .  ::  :.        .::::  :::: : . *  . .:  ::.. :

Tt.Ymf60   FNNKINLFNNEFN----KFIYSWDNFFFNKIKFTGKGFKWKKKETNLFLFFNRAH-KCFFIGNN
Tm.Rpl6    WR-KVKMFQGTYWSLIRNMVVGVTEGYKKELEIVGIGYRAQLQGNTLVMNLGYAHPVVYEIPSD
           :.  *:::*:. :     ::: .  : : :::::.* *:: : : ..*.: :. **   : * .:

Tt.Ymf60   IILKRLSKSKVILLKNNYNHLIHDSILVRTIRSNNIFTKRGLRFSRQIILKKKGKTASQ
Tm.Rpl6    VKIEVPAPNRIIVSGIDKQRVGQVAAEIRAFRPPNVYTGKGIRYVGEVVRQKEGKKA--
           : ::  : .::*:   : ::: : :  :*::*. *::* :*:*:  ::: :*:**.*

b
Tt.Ymf62   MFLITITSYFSNIIEFNSYIINLIDFITPLFFIENFVIQFFILYLFYLLIVNNNLYYILLYIFL
Pp.Nad6    ---------MTSITLF--WIFSIISLLSACLVVSLANAVYSVLFLIIVFCNTASILLLLGAEFL
                    ::.*  *  :*:.:*.:::. :.:.       : :*:*: ::  . .:  :*   **

Tt.Ymf62   EIVFFGLFLCLYQLELFTGFLWVAEFAIVFIAVVLLFYLNIDGLHLKYNHNINNVLYYTPSLVL
Pp.Nad6    SFLFL---------------IVYVGAIAVLFLFVVMMLNVKIDGAKVNYTT-ILLIGFLISSILL
           .::*:               :::*. :*::*: **::: ::*** :::*.  *  : :  .*::*

Tt.Ymf62   FLIFFNIDYFSELELFLPLELSFIDIYDDYYEGFNNSIMNDFTPLTLSYYSINSAEFIIIGLLL
Pp.Nad6    FQFSSAMQIDLDAYSNIRLNLTEVNFPTTTLWTQENSLPSNVESIGLVLYTSYSLVFILCAFIL
           * :   ::   :    : *:*: :::           :**: .:. .: *  *:  *  **: .::*

Tt.Ymf62   LLGSVACVNLYKSNKNYTIVKQSNLLTMFDFFKDFINFSFIRKQDLNNQTNFNPSLRSIKKKY
Pp.Nad6    LLAMIGSIILTMNQRN-GVKKQQITLQLHRDQNKVVRFLDLRKN-------------------
           **. :..: *  .::*  : **.  * :.   :..:.*  :**:

c
Tt.Ymf58   --MLTWISFWSLIFWLILIILVLKPKNFISILFMSELTWLALYCLSLLFGAIYCDITLLS
No.Nad4L   MNQINFLAVSMILFLLGIWGIFLNRKNIIIMLMSIELMLLAVNLNFLFFSVYLDDMIGQI
              :.:::.  ::* * :  :.*: **:* :*:  **  **:    *:*..   *:

Tt.Ymf58   ISFFILGVAGLEFSFGILIAILYKNLNESLNTDLNNNNNNQNIFDKNFKTPLEKINWQ
No.Nad4L   FALFILTVAAAESAIGLALLVSYYRVRGSIAVELIHLMKG------------------
           :::*** **. * ::*: : : * .:. *: .:* :  :.
```

Figure 2. Protein sequence alignments for (**a**) *T.thermophila* Ymf60 with *T.maritima* Rpl6, (**b**) *T.thermophila* Ymf62 with *P.monodon* Nad6 and (**c**) *T.thermophila* Ymf58 with *N.olivacea* Nad4L. * indicates identical amino acids; : indicates highly similar amino acids; . indicates similar amino acids as designated by the Clustal X program.

On the basis of the additional physico-chemical parameters, Ym62 and Ymf58 appear to be proteins involved in oxidative phosphorylation and electron transport. They have five and three TM respectively and they each have low pIs, ratios of negative to positive residues above 1.0, aliphatic indices well above 100 and positive GRAVYs. The TM profiles of Ymf62 and Ymf58 were compared with the TM profiles of representatives of all of the proteins involved in oxidative phosphorylation and electron transport whose function had not already been identified within the *T.thermophila* mitochondrial genome.

ORF Ymf62 has a TM profile very similar to that of Nad6 from *B.taurus* (Fig. 3A). Each sequence has four TM helices followed by a non-TM region of ~40 amino acids followed by a fifth TM region. A survey of virtually all of the Nad6 proteins in GenBank indicates that the Nad6 protein of *Porphyra purpurea* has the greatest sequence similarity to the *T.thermophila* Ymf62 protein (Fig. 2b). A chain of homology was established linking Ymf62 to *B.taurus* Nad6 (Table 2). The combination of TM profile and sequence similarity allows us to assign Nad6 function to the *T.thermophila* Ymf62 protein.

The TM profile of Ymf58 is very similar to the TM profile of *B.taurus* Nad4L, both for the TM regions and the inter-TM spacings (Fig. 3B). A survey of virtually all of the Nad4L proteins in GenBank indicates that Nad4L of *Nephroselmis olivacea* has the greatest sequence similarity to the *T.thermophila* Ymf58 protein (Fig. 2c). Although the sequence similarity is only moderate (5.47), coupled with



Tetrahymena thermophila Ymf62

Tetrahymena thermophila Ymf58

Bos taurus Nad6

Bos taurus Nad4L

Figure 3. TM profiles for (**A**) *T.thermophila* Ymf62 and *B.taurus* Nad6 and the TM profiles for (**B**) *T.thermophila* Ymf58 and *B.taurus* Nad4L. Horizontal lines at the top of the figures indicate 'inside' or 'outside' and blocks represent TM regions.

the similarity of TM profiles, homology is highly probable. A chain of homology extending from *T.thermophila* to the *B.taurus* Nad4L sequence via *N.olivacea* and *Penaeus monodon* can be readily established (Table 2). The function of Nad4L can be assigned to the *T.thermophila* Ymf58.

The ORF identified as *yejR* by Burger *et al.* (7) does not have an apparent homolog in either *S.cerevisiae* or *B.taurus*. However, this ORF is clearly homologous to a *Triticum aestivum* (wheat) gene involved in the biogenesis of c-type cytochromes via a homology chain through *Pseudomonas putida* (Table 2). The *T.aestivum* protein is a member of a family that consists of various proteins involved in cytochrome *c* assembly from mitochondria and bacteria (41). Thus, the *T.thermophila* YejR protein is assigned a similar function. This protein family differs from the heme lyase found in yeast (42).

Virtually all of the identified ORFs have reasonably strong hits with the appropriate domains in the PRODOM and pFAM databases. This strengthens our confidence that the appropriate function has been assigned to these ORFs. Most of the ORFs without assigned functions do not have any hits with domains in the PRODOM and pFAM databases. The few hits that do occur for these ORFs have not led to the identification of proteins with reasonably high sequence similarity.

One of the *T.pyriformis* ORFs, which was identified as Rps3 by Burger *et al.* (7), does not have significant sequence similarity with any of the Rps3 genes found in GenBank. No protein has been found that has significant sequence similarity to this ORF. Most of the physico-chemical parameters suggest that this protein is potentially a ribosomal protein. However, it has a predicted TM, which is not characteristic of ribosomal proteins (Table 4).

The *T.thermophila* mitochondrial genome has 19 ORFs, in addition to Rps3, that do not have sufficient sequence similarity to proteins of known function to permit the confident assignment of function. On the basis of their physico-chemical parameters these ORFs can be grouped into two broad classes: putative ribosomal proteins and putative non-ribosomal proteins. Table 4 lists the parameters of these ORFs along with the mean values of these parameters from the ORFs for which function has been assigned.

DISCUSSION

The pattern of nucleotide substitution between homologous genes gives a good indication of the selective pressure on a coding region. In general, mitochondrial proteins for which function can be assigned have higher Ks/Ka ratios than proteins of unknown function, whereas proteins involved in oxidative phosphorylation have among the highest Ks/Ka ratios. Genes *atp9* and *nad5* are exceptions. Their Ks/Ka ratios indicate less selective pressure than expected for proteins with assigned function. The divergence at the C-terminus of *atp9* may account for its low Ks/Ka ratio, but the reason for the low

Table 4. Physico-chemical properties and TM predictions for unidentifiable mitochondrial ORFs

	Gene	Accession no.	Theoretical pI	Ratio of –/+ residues	Aliphatic index	GRAVY	TM
Putative ribosomal	*ymf59*	NP_149385	10.23	0.174	98.09	–0.103	0
	ymf61	NP_149388	10.14	0.275	93.78	–0.429	0
	ymf63	NP_149382	9.71	0.435	90.72	–0.349	0
	ymf64	NP_149378	10.18	0.286	98.03	–0.305	0
	ymf76	NP_149366	10.87	0.113	92.20	–0.554	0
	'*rps3*'	NP_149368	10.14	0.167	85.83	–0.340	1
			10.58	0.318	84.04	–0.510	0
Putative non-ribosomal	*ymf56*	NP_149398	6.92	1.00	109.48	0.198	1
	ymf57	NP_149364	9.78	0.461	109.01	0.646	2
	ymf65	NP_149383	9.32	0.450	137.12	1.017	10
	ymf66	NP_149365	4.88	1.792	103.97	0.462	8
	ymf67	NP_149399	9.79	0.364	115.32	0.061	5
	ymf68	NP_149400	9.45	0.610	109.29	0.211	8
	ymf69	NP_149384	10.13	0.231	112.36	0.242	1
	ymf70	NP_149406	9.64	0.333	121.69	0.479	1
	ymf71	NP_149401	9.22	0.333	168.91	1.546	3
	ymf72	NP_149390	7.69	0.875	135.13	0.532	3
	ymf73	NP_149408	9.88	0.333	114.03	0.111	0
	ymf74	NP_149371	9.88	0.346	126.12	0.077	0
	ymf75	NP_149379	9.74	0.346	118.42	0.403	2
	ymf77	NP_149394	9.82	0.384	129.47	0.124	15
			7.47	1.08	113.31	0.490	4.8

Unidentified genes can be classified as ribosomal or non-ribosomal based on these trends. Bold numbers are mean values transferred from Table 2 for identified ribosomal (top) and non-ribosomal proteins (bottom).

Ks/Ka ratio of *nad5* is unclear. The proteins for which function cannot be readily assigned have generally lower Ks/Ka ratios, with some exceptions like Ymf56, Ymf65 and Ymf68, which have Ks/Ka ratios >10. The largest ORF, *ymf77*, has a very low Ks/Ka ratio (1.61) suggesting reduced selective pressure, which may contribute to the divergence of this protein and the difficulty in matching it with a protein of known function.

The mitochondrial genome is very compact with very short intergenic regions (3.8% of the genome), only three of which are longer than 63 bp. These intergenic regions appear to be too small to accommodate promoters. This is consistent with the transcript mapping of the *T.pyriformis* mitochondrial genes which suggested multi-gene transcripts (16). A working hypothesis is that the entire genome is transcribed from a central bi-directional promoter, which would account for the transcription of all of the genes except the three gene cluster in the 5′ portion of the genome. These three genes could be transcribed from a unidirectional promoter in the 5′→3′ direction (Fig. 1).

The three long intergenic regions in *T.thermophila* have sequence *z* scores indicating that these regions are under strong selective pressures (data not shown). The largest intergenic region (493 bp), between *ymf77* and *cob*, corresponds to the site of the putative central bi-directional promoter set. This position also corresponds to the origin of DNA replication identified in electron micrographs of partially replicated mitochondrial molecules in *T.pyriformis* (43). The sequence of the second largest intergenic region (261 bp) is highly conserved relative to the *T.pyriformis* homolog. It is located upstream from and immediately adjacent to the three gene cluster transcribed in the 5′→3′ direction. This may be the site of a unidirectional promoter responsible for transcribing this three gene cluster.

The most notable difference between the *T.thermophila* and *T.pyriformis* mitochondrial genomes is the tandem duplication of the *nad9* gene in *T.thermophila*. A similar duplication occurs in *T.malaccensis* and the divergence between the genes from different species (orthologous) is much greater than the divergence among genes within the species (paralogous) indicating concerted evolution (Table 3a). Concerted evolution also appears to be operating on the terminally repeated regions (Table 3b). The regions under concerted evolution include the two split portions of the *LSU rRNA* gene, the leucine *tRNA* gene interrupting the *LSU rRNA* sequence and the two short (3 and 9 bp) intergenic regions flanking the tRNA gene. The telomeres at both ends of the mitochondrial DNA molecule (53 bp repeats) are also identical although they start at different positions in the telomere repeat sequence. Morin and Cech have proposed that unequal crossing over and terminal hybridization may be part of a process maintaining the telomeres (11,44). Whether this is the mechanism by which telomeres are regenerated or not, it is not a plausible mechanism for the concerted evolution of the inverted terminal repeats as it would be expected to homogenize the terminal tRNA genes. Concerted evolution of paralogs is relatively common (45–47).

The sequences for the C-terminus of the Atp9 protein and the intergenic region following the *atp9* gene from *T.thermophila* and *T.pyriformis* have diverged significantly. The *T.malaccensis* and *T.thermophila atp9* gene sequences are very similar at the C-termini and they are more similar to the Atp9 proteins from other organisms than is the *T.pyriformis* sequence. The most probable evolutionary history is that a DNA sequence was inserted into the C-terminus of the *atp9* gene in the lineage leading to *T.pyriformis*, creating a slightly different C-terminal sequence and substantially lengthening the following intergenic region.

One of the putative mitochondrial proteins in *Tetrahymena*, Ymf77, presents a particular challenge. This gene has 1321 amino acids (almost 9% of the *Tetrahymena* mitochondrial genome) and apparently has 15 TM. It is conserved in both *T.thermophila* and *T.pyriformis*, although not found in

Paramecium aurelia (48). In spite of its size and apparent function, BLASTP does not yield any reasonable protein candidates. Additional mitochondrial sequences from ciliates intermediate between *Tetrahymena* and *Paramecium* may offer additional versions of Ymf77 that give clues to its function.

The proteins found in mitochodrial genomes are relatively limited in potential function, which simplifies the identification of mitochondrial ORFs (13). In most mitochondrial genomes all of the genes are identified, even *R.americana* with 97 mitochondrial genes has virtually all of its genes functionally identified (22). In spite of this, almost half of the putative proteins in *Tetrahymena* remain unidentified. Clearly the mitochondrial genomes of the ciliates have diverged dramatically from other mitochondrial genomes, however, the evolutionary mechanisms responsible for this divergence are not immediately obvious.

Although sequence similarity remains the primary means of establishing homology from which function can be assigned, physico-chemical parameters provide a powerful additional indicator of potential homology. These parameters can be combined with nominal sequence similarity to establish homology. They are also valuable in clustering unidentified proteins into specific groups.

ACKNOWLEDGEMENTS

We wish to thank Dr Eduardo Orias for providing *T.thermophila* strain SB210 and Dr Howard Buhse for providing *T.pyriformis* strain GL. Appreciation is expressed to Dr Erik Avaniss-Aghajani and Dr Kenneth Jones for helpful suggestions and critical advice. This work was supported in part by a grant from the Academic Senate of the University of California at Los Angeles.

REFERENCES

1. Suyama,Y. and Miura,K. (1968) Size and structural variation of mitochondrial DNA. *Proc. Natl Acad. Sci. USA.*, **60**, 235–242.
2. Brunk,C.F. and Hanawalt,P.C. (1969) Mitochondrial DNA in *Tetrahymena pyriformis*. *Exp. Cell Res.*, **554**, 143–149.
3. Chiu,N., Chiu,A. and Suyama,Y. (1975) Native and imported transfer RNA in mitochondria. *J. Mol. Biol.*, **99**, 37–50.
4. Suyama,Y. (1982) Native and imported tRNAs in *Tetrahymena* mitochondria: evidence for their involvement in intramitochondrial translation. In Slonimski,P., Borst,P. and Attardi,G. (eds), *Mitochondrial Genes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 449–455.
5. Rusconi,C.P. and Cech,T.R. (1996) The anticodon is the signal sequence for mitochondrial import of glutamine tRNA in *Tetrahymena*. *Genes Dev.*, **10**, 2870–2880.
6. Rusconi,C.P. and Cech,T.R. (1996) Mitochondrial import of only one of three nuclear-encoded glutamine tRNAs in *Tetrahymena thermophila*. *EMBO J.*, **15**, 3286–3295.
7. Burger,G., Zhu,Y., Littlejohn,T.G., Greenwood,S.J., Schnare,M.N., Lang,B.F. and Gray,M.W. (2000) Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. *J. Mol. Biol.*, **297**, 365–380.
8. Arnberg,A.C., van Bruggen,E.F.J, Borst,P., Clegg,R.A., Schutgens,R.B.H., Weijers,P.J. and Goldbach,R.W. (1975) Mitochondrial DNA of *Tetrahymena pryiformis* strain ST contains a long terminal duplication-inversion. *Biochim. Biophys. Acta*, **383**, 359–369.
9. Morin,G.B. and Cech,T.R. (1988) Mitochondrial telomeres: surprising diversity of repeat telomeric DNA sequences among six species of *Tetrahymena*. *Cell*, **52**, 367–374.
10. Heinonen,T.Y.K., Schnare,M.N., Young,P.G. and Gray,M.W. (1987) Rearranged coding segments, separated by transfer RNA genes, specify the two parts of a discontinuous large subunit ribosomal RNA in *Tetrahymena pyriformis* mitochondrial DNA. *J. Biol. Chem.*, **262**, 2879–2887.
11. Morin,G.B. and Cech,T.R. (1988) Phylogenetic relationships and altered genome structures among *Tetrahymena* mitochondrial DNAs. *Nucleic Acids Res.*, **16**, 327–347.
12. Brunk,C.F., Tran,A.B., Lee,L.C. and J. Li. (2001) Complete sequence of the mitochondrial genome of *Tetrahymena thermophila* and comparison with the mitochondrial genome of *Tetrahymena pyriformis.* GenBank accession no. AF396436.
13. Gray,M.W., Lang,B.F., Cedergren,R., Golding,G.B., Lemieux,C., Sankoff,D., Turmel,M., Brossard,N., Delage,R., Littlejohn,T.G., Plante,I., Rioux,P., Saint-Louis,D., Zhu,Y. and Burger,G. (1998) Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.*, **26**, 865–878.
14. Sadler,L.A. and Brunk,C.F. (1992) Phylogenetic relationships and unusual diversity in histone H4 proteins within the *Tetrahymena pyriformis* complex. *Mol. Biol. Evol.*, **9**, 70–84.
15. Nanney,D.L., Park,C., Preparata,R. and Simon,E.M. (1998) Comparison of sequence differences in a variable 23S rRNA domain among sets of cryptic species of ciliated protozoa. *J. Eukaryot. Microbiol.*, **45**, 91–100.
16. Edqvist,J., Burger,G. and Gray,M.W. (2000) Expression of mitochondrial protein-coding genes in *Tetrahymena pyriformis*. *J. Mol. Biol.*, **297**, 381–393.
17. Gray,M.W. (1992) The endosymbiont hypothesis revisited. *Int. Rev. Cytol.*, **141**, 233–257.
18. Adams,K.L., Daley,D.O., Qiu,Y.L., Whelan,J. and Palmer,J.D. (2000) Repeated, recent and diverse transfer of a mitochondrial gene to the nucleus in flowering plants. *Nature*, **408**, 354–357.
19. Germot,A., Philippe,H. and Le Guyader,H. (1996) Presence of a mitochondrial-type 70-kDa heat shock protein in *Trichomonas vaginalis* suggests a very early mitochondrial endosymbiosis in eukaryotes. *Proc. Natl Acad. Sci. USA*, **93**, 14614–14617.
20. Hashimoto,T., Sanchez,I.B., Shirakura,T., Muller,M. and Hasegawa,M. (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc. Natl Acad. Sci. USA*, **95**, 6860–6865.
21. Gray,M.W., Burger,G. and Lang,B.F. (1999) Mitochondrial evolution. *Science*, **283**, 1476–1481.
22. Lang,B.F., Burger,G., O'Kelley,C.J., Cedergren,R., Golding,G.B., Lemieux,C., Sankoff,D., Turmel,M. and Gray,M.W. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature*, **387**, 493–497.
23. Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.
24. Orias,E. and Bruns,P.J. (1976) Induction and isolation of mutants in *Tetrahymena*. In Prescott,D.M. (ed.), *Methods in Cell Biol.* Academic Press, New York, NY, Vol. 13, pp. 247–283.
25. Nanney,D.L. and Simon,E.M. (2000) Laboratory and evolutionary history of *Tetrahymena thermophila*. In Asai,D.J. and Forney,J.D. (eds), *Tetrahymena thermophila*. Academic Press, San Diego, CA, pp. 4–25.
26. Thorstenson,Y.R., Hunicke-Smith,S.P., Oefner,P.J. and Davis,R.W. (1998) An automated hydrodynamic process for controlled, unbiased DNA shearing. *Genome Methods*, **8**, 848–855.
27. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
28. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using Phred II. Error probabilities. *Genome Res.*, **8**, 186–194.
29. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
30. Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
31. Higgins,D.G. and Sharp,P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Comput. Appl. Biosci.*, **5**, 151–153.
32. Doolittle,R.F. (1987) *Of URFS and ORFS: A Primer on How to Analyze Dervied Amino Acid Sequences*. University Science Books, Mill Valley, CA.
33. Doolittle,R.F. (1990) Searching through sequence databases. *Methods Enzymol.*, **183**, 99–110.

34. Mount,D.W. (2001) *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 96–118

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

36. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.

37. Comeron,J.M. (1995) A method for estimating the number of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.*, **41**, 1152–1159.

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

39. Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.*, **19**, 258–260.

40. Sonnhammer,E.L.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting TM helices in protein sequences. In Glasgow,J., Littlejohn,T., Major,F., Lathrop,R., Sankoff,D. and Sensen,C. (eds), *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 175–182.

41. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

42. Dumont,M., Ernst,J., Hampsey,D. and Sherman,F. (1987) Identification and sequence of the gene encoding cytochrome C heme lyase in the yeast *Saccharomyces cerevisiae*. *EMBO J.*, **6**, 235–241.

43. Arnberg,A.C., van Bruggen,E.F.J., Clegg,R.A., Upholt,W.B. and Borst,P. (1975) An analysis by electron microscopy of intermediates in the replication of linear *Tetrahymena* mitochondrial DNA. *Biochim. Biophys. Acta*, **361**, 266–276.

44. Morin,G.B. and Cech,T.R. (1986) The telomeres of the linear mitochondrial DNA of *Tetrahymena thermophila* consist of 53 bp tandem repeats. *Cell*, **46**, 873–883.

45. Brown,D.D, Wensink,P.C. and Jordan,E. (1972) A comparison of the ribosomal DNAs of *Xenopus laevis* and *Xenopus mulleri*: evolution of tandem genes. *J. Mol. Biol.*, **63**, 57–73.

46. Dover,G.A. (1993) Evolution of genetic redundancy for advanced players. *Curr. Opin. Genet. Dev.*, **3**, 902–910.

47. Elder,J.F. and Turner,B.J. (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. *Q. Rev. Biol.*, **70**, 279–320.

48. Prichard,A.E., Seilhamer,J.J., Mahalingam,R., Sable,C.L., Venuti,S.E. and Cummings,D.J. (1990) Nucleotide sequence of the mitochondrial genome of *Paramecium. Nucleic Acids Res.*, **18**, 173–180.