Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 229–234, January 1998
Evolution

A mitochondrial-like chaperonin 60 gene in *Giardia lamblia:* Evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria

ANDREW J. ROGER†, STAFFAN G. SVÄRD‡, JORGE TOVAR§, C. GRAHAM CLARK§, MICHAEL W. SMITH¶, FRANCES D. GILLIN‡, AND MITCHELL L. SOGIN†‖

†The Josephine Bay–Paul Center of Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543-1015; ‡Department of Pathology, University of California at San Diego Medical Center, San Diego, CA 92103-8416; §Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom; and ¶Science Applications International Corporation Frederick, National Cancer Institute, Frederick, MD 21702-1201

Communicated by David M. Prescott, University of Colorado, Boulder, CO, and approved November 17, 1997 (received for review October 7, 1997)

ABSTRACT **Diplomonads, parabasalids, as represented by trichomonads, and microsporidia are three protist lineages lacking mitochondria that branch earlier than all other eukaryotes in small subunit rRNA and elongation factor phylogenies. The absence of mitochondria and plastids in these organisms suggested that they diverged before the origin of these organelles. However, recent discoveries of mitochondrial-like heat shock protein 70 and/or chaperonin 60 (cpn60) genes in trichomonads and microsporidia imply that the ancestors of these two groups once harbored mitochondria or their endosymbiotic progenitors. In this report, we describe a mitochondrial-like cpn60 homolog from the diplomonad parasite *Giardia lamblia*. Northern and Western blots reveal that the expression of cpn60 is independent of cellular stress and, except during excystation, occurs throughout the *G. lamblia* life cycle. Phylogenetic analyses position the *G. lamblia* cpn60 in a clade that includes mitochondrial and hydrogenosomal cpn60 proteins. The most parsimonious interpretation of these data is that the cpn60 gene was transferred from the endosymbiotic ancestors of mitochondria to the nucleus early in eukaryotic evolution, before the divergence of the diplomonads and trichomonads from other extant eukaryotic lineages. A more complicated explanation requires that these genes originated from distinct α-proteobacterial endosymbioses that formed transiently within these protist lineages.**

The diplomonad protist *Giardia lamblia*, a principal cause of diarrheal disease (1), is basal to all eukaryotes with mitochondria in phylogenies inferred from small subunit rRNAs (2, 3) and several protein-coding genes (4, 5). The early emergence of diplomonads, trichomonads, and microsporidia in molecular trees, coupled with their lack of mitochondria, supports the view that these organisms diverged before the endosymbiotic origin of mitochondria within eukaryotes (2, 5–8). However, discoveries of mitochondrial-like chaperonin 60 (cpn60), chaperonin 10 (cpn10), and heat shock protein 70 (hsp70) genes in a trichomonad (9–12) and hsp70 homologs in microsporidia (13, 14) challenge this interpretation. Mitochondrial cpn60, cpn10, and hsp70 proteins are encoded by nuclear genes that are specifically related to homologs in α-Proteobacteria, indicating that they are of endosymbiotic origin (9, 11, 15). The presence of mitochondrial-like homologs of these genes in trichomonads and microsporidia implies that the ancestors of these organisms once harbored mitochondria, or their endosymbiotic progenitors.

However, evidence for loss of mitochondrial functions from the diplomonad lineage is more tenuous (16). Phylogenies of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and triosephosphate isomerase (TPI) are poorly resolved but do show that eukaryotic homologs could have entered the eukaryotic nucleus by transfer from the ancestral mitochondrial endosymbiont (17, 18). If this interpretation is correct, the presence of typical eukaryotic GAPDH and TPI genes in *G. lamblia* suggests mitochondria or their ancestors were lost from the diplomonad lineage (17, 18). An ancestral mitochondrial endosymbiont in diplomonads was also supported by the report of a 60-kDa protein from *G. lamblia* that cross-reacts with mammalian mitochondrial cpn60 antibodies (19). Yet, none of these examples establishes a specific link with the mitochondrial lineage. Independent lateral transfer of genes from prokaryotes to eukaryotes (20) could explain the GAPDH and TPI phylogenies as well as the immunological cross-reactivity data.

Here we report the isolation and sequence analysis of a cpn60 gene from *G. lamblia* that is phylogenetically related to the mitochondrial cpn60 lineage. These data suggest that the α-proteobacterial endosymbiont that gave rise to mitochondria may have entered the eukaryotic lineage much earlier than previously thought, possibly before the divergence of all known eukaryotes.

MATERIALS AND METHODS

Cloning and Sequencing of *G. lamblia* cpn60. Basic local alignment tool (BLASTX) searches of sequences from randomly selected *G. lamblia* cosmids identified the presence of a partial mitochondrial-like cpn60 gene (21). An ≈850-bp fragment from the end of cosmid CLM-8f8 was subcloned into the pBluescript plasmid vector (Stratagene) and used as a probe to screen a *G. lamblia* λZAPII genomic library (22) by using digoxigenin-labeling and detection methods (Boehringer Mannheim). After secondary screening, *in vivo* excision converted the positive clones into pBluescript plasmids (Stratagene). We determined the sequences of the *G. lamblia* cpn60 homolog as well as its immediate upstream and downstream regions on both DNA strands.

Cycle sequencing reactions were carried out on all plasmid and cosmid genomic clones by using the Sequitherm Long-read and Excel II kits (Epicentre Technologies, Madison, WI)

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: cpn, chaperonin; hsp, heat shock protein; HKY, Hasegawa–Kishino–Yano.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF029695 and AF029366).

‖To whom reprint requests should be addressed. e-mail: sogin@evol5.mbl.edu.

with dye-labeled M13 forward, M13 reverse, T3, and T7 primers. Reactions were run on a LI-COR 4200 automated sequencer, and sequence data were collected and edited by using LI-COR software (LI-COR, Lincoln, NE).

We also determined the full-length sequence of *Entamoeba histolytica* cpn60, completing the partial sequence previously reported (15).

Growth of Cells. *G. lamblia* strain WB (ATCC 30957), clone C6 trophozoites were grown, encysted, and excysted as previously described (23).

Preparation of RNA and DNA. Total RNA was isolated from *G. lamblia* at the various stages of differentiation by extraction with RNazol B (Tel-Test, Friendswood, TX). Genomic DNA was isolated by using the Qiagen Blood and Culture DNA Kit (Qiagen, Chatsworth, CA).

Southern and Northern Analyses. The probe used for Southern and Northern blots was a random primer-labeled PCR fragment amplified from *G. lamblia* genomic DNA with the *G. lamblia*-specific oligonucleotides GlcpnR1 (5′-AACCGGACAGGTTCATGAAG-3′) and GlcpnF1 (5′-ATAGGCAGACTATTGGGTAAG-3′). Southern hybridization was performed as described previously (22). For Northern hybridization, 20 μg of total RNA was fractionated in 1.5% formaldehyde-agarose gels, capillary blotted, and immobilized onto nylon membranes (Zeta-Probe, Bio-Rad). The membranes were prehybridized, hybridized and washed at high-stringency, and autoradiographed by using standard techniques.

Rapid Amplification of cDNA Ends (RACE) Analyses. 5′-RACE and 3′-RACE techniques (System 2; GIBCO/BRL) identified the starts of transcription and the polyadenylation site of cpn60. Oligonucleotide GlcpnR1 primed the first strand, and oligonucleotide Glcpn-2 (5′-GAGCAGC-CCGGGGCTGCAGAGAGAG-3′) served as a nested primer. 3′-RACE was performed on cDNA generated from 10 μg total RNA by using Superscript II (GIBCO/BRL), using poly-T oligo SGS-10 (5′-CGAGCTGCGTCGACAGGC$(T)_{17}$-3′) and gene-specific oligo GlcnpF1. PCR conditions were 94°C, 1 min; 55°C, 1 min; 72°C, 1 min for 30 cycles. For the production of DNA sequencing templates, the PCR products were cloned into the pGEM-T EASY vector (Promega).

Western Blot Analyses. Western blots from encysting and excysting cells were prepared as described in ref. 24. Blots were reacted with rabbit anti-cpn60 (diluted 1:250) antibodies raised against *Synechococcus* sp. GroEL (StressGen Biotechnologies, Victoria, Canada) and probed with protein A-alkaline phosphatase conjugate. Controls for equal loading were reacted with monoclonal antibodies (diluted 1:250) to the *G. lamblia* lectin, taglin (25), and rabbit polyclonal antibodies to the endoplasmic reticulum protein BiP (26).

Stresses. Attached trophozoites or 18-hr encysting cells were subjected to heat shock (40°C or 43°C) for 20 min, then allowed to recover at 37°C for 60 or 90 min. Encysting cells were also incubated in 3% ethanol for 20 min or DTT (7.5 mM) for 3 hr and allowed to recover for 0, 60, or 90 min.

Electron Microscopy. Cells were harvested at the indicated times, and pellets were fixed and processed for cryosection immunoelectron microscopy, as described in ref. 27, then reacted with the *Synechococcus* sp. anti-cpn60 antibody followed by localization with 5 nm gold-labeled goat anti-rabbit antibodies.

Sequence Alignment. A database containing 121 eubacterial GroEL, plastid and mitochondrial cpn60 homologs, archaebacterial thermophilic factor (tf) homologs, and eukaryotic t-complex polypeptide-1 (tcp) homologs was assembled from GenBank and Swiss-Prot databases. Sequences were aligned by using the CLUSTALW program (28). Adjustments to the alignment were made by using the SEQLAB program (GCG), and regions of ambiguous alignment were removed to create datasets for phylogenetic analysis. The first dataset contained 179 aligned amino acid positions from 47 sequences representing all three domains of life. A second dataset contained 513 aligned amino acid positions and included only mitochondrial and eubacterial cpn60 homologs.

Protein Phylogeny. Protein distance matrices were inferred by using the PROTDIST program (Dayhoff PAM model), and trees were generated by using the FITCH program with global rearrangements (29). Unweighted maximum parsimony analysis was carried out by 50 rounds of random stepwise addition heuristic searches with tree bisection reconnection (TBR) branch swapping, by using the PAUP* 4.0d56 program (30). Protein maximum likelihood (ML) trees were inferred by using the protein maximum likelihood program PROTML 2.2 (31) and the heuristic quick-add OTU searching method with the Jones, Taylor, and Thornton (JTT-f) amino acid replacement model. Deviations of amino acid frequencies in a given sequence from the overall frequencies in the dataset and estimates of the gamma shape parameter, α (describing the degree of rate variation among sites in the dataset), were evaluated by using the PUZZLE 3.1 program (32).

Nucleotide Phylogeny. All nucleotide phylogenetic analyses were carried out by using the PAUP* 4.0d56 program (30). Nucleotide distance matrices were inferred from the first and second codon positions of corresponding nucleotide alignments by the maximum likelihood distance method employing the Hasegawa–Kishino–Yano substitution model with a four-category discrete approximation to the gamma distribution (HKY+ Γ). Maximum likelihood estimates of the transition/transversion (Ti/Tv) ratio and the gamma shape parameter (α) were based on the nucleotide distance/neighbor-joining topology. Trees were inferred by simple stepwise-addition heuristic searches with TBR branch swapping under the minimum evolution optimality criterion. Nucleotide maximum likelihood analysis was performed by using the HKY+Γ model and tree-searching methods as described above.

Bootstrap Analyses. Bootstrap analyses for protein distance analysis utilized programs of the PHYLIP package (29). The resampling estimated log-likelihood (RELL) method was used to estimate bootstrap values for protein maximum likelihood analyses (31). All other bootstrap analyses were carried out by using the PAUP* 4.0d56 program (30). Distance and parsimony analyses were based on 500 bootstrap resamplings whereas nucleotide maximum likelihood analysis was based on 250 resamplings.

RESULTS AND DISCUSSION

Characterization of the *G. lamblia* cpn60 Gene. Of approximately 2,600 single-pass sequences from the ends of *G. lamblia* cosmids (21), two displayed significant similarity to the GroEL/cpn60 gene family. After secondary screening of a *G. lamblia* λZAPII genomic library (22), positive clones were subcloned and sequenced on both DNA strands, yielding a total of 2,777 bp containing the cpn60 gene and flanking regions.

The sequence of the ORF (547 codons) shows clear homology to both eubacterial GroEL and mitochondrial cpn60 over its entire length. No other ORFs were detected in the immediate upstream or downstream regions. Southern blots were consistent with a single-copy cpn60 gene in the *G. lamblia* genome.

The sequence of 5′-RACE products indicated that the start site of the *G. lamblia* cpn60 gene lies at positions −5 and −4 upstream of the start codon (Fig. 1), well within the range of 1–11 nt observed for other *G. lamblia* 5′ untranslated regions (1, 22, 33). An AT-rich motif, corresponding to the 8- to 9-bp AT-rich initiator sites reported in other *G. lamblia* genes (33, 34), extends from positions −8 to +1 (Fig. 1). A second upstream motif, AAATTT, spans positions −46 to −41 (Fig.

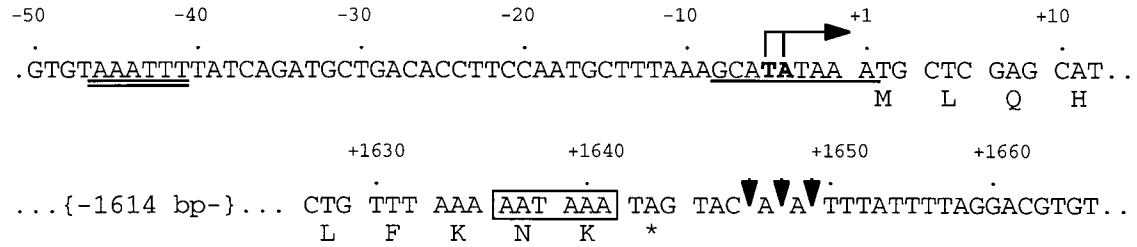

FIG. 1. Properties of the *G. lamblia* cpn60 gene and mRNA transcript. Upstream and downstream regions of the cpn60 gene are shown and numbered relative to the first base of the start codon (+1). The transcriptional start sites are shown in bold under vertical lines and rightward-pointing arrow (→). Two possible transcriptional signals are identified: an AT-rich transcription initiation signal (single underline), and an upstream promoter element (double underline) similar to the CAATTT signal reported for other *G. lamblia* genes (33). Possible sites of polyadenylation (↓), a putative polyadenylation signal (boxed), and the stop codon (*) are also indicated.

1) and resembles the 6-base consensus motif CAATTT present upstream of other *G. lamblia* coding regions (33).

Sequencing of 3′-RACE products revealed the presence of a poly(A) tail on the *G. lamblia* cpn60 transcript beginning 4 nt downstream of the stop codon (Fig. 1). A 6-nt motif, AATAAA, has only a single mismatch with the putative consensus polyadenylation signal AGTRAA for *G. lamblia* genes (1, 22) and it immediately precedes the stop codon (Fig. 1).

Expression of the cpn60 Gene Throughout *G. lamblia* Life Cycle. Northern analysis of total *G. lamblia* RNA showed a single band of approximately 1.8 kb in length that strongly hybridized with the cpn60 probe (Fig. 2*A*). Because changes in expression of molecular chaperones are observed during differentiation of certain parasites (35), we followed the expression of cpn60 throughout the *G. lamblia* life cycle.

First, we determined whether the expression of the transcript was affected by differentiation from the vegetative trophozoite to cyst life-cycle stages in response to elevated pH and bile concentrations, which induce encystation. Levels of cpn60 transcript are nearly constant during the *in vitro* encystation process, with a slight decrease in expression late in encystation (48 hr) (Fig. 2*A*), mirroring the decrease in levels of many transcripts observed at this stage in the *G. lamblia* life cycle (22).

Western blots were prepared from total *G. lamblia* protein and probed with an anti-cpn60 antibody raised against *Synechococcus* sp. GroEL. The antibody reacted with a band in the 60-kDa range, corresponding to the expected size of the cpn60 gene product (Fig. 2*B*). The level of this protein did not change appreciably during the encystation process (Fig. 2*B*).

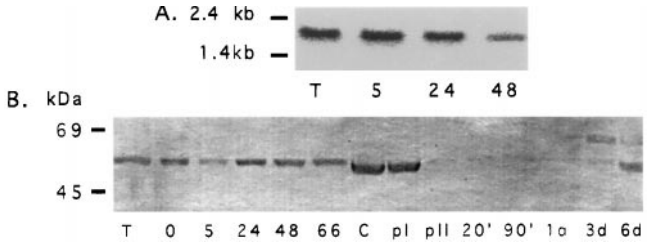

FIG. 2. Expression of cpn60 at different stages in the *G. lamblia* life cycle. (*A*) Northern blot analysis of total cellular RNA using a PCR fragment of the *G lamblia* cpn60 as a probe. Expression of the ≈1.8-kb cpn60 transcript was monitored in the vegetative trophozoite (T) stage, and after 5, 24, and 48 hr of encystation of *G. lamblia.* (*B*) Western blot analysis of total cellular protein isolated at different stages of encystation and excysation using a *Synechococcus* sp. anti-cpn60 antibody. Levels of the ≈60-kDa putative cpn60 protein were constant during the transition from the trophozoite stage (T) and throughout 0–66 hr of encystation. This protein was also present in the cyst phase (C) and after the first stage of induction of excystation (pI). A marked decrease in cpn60 protein was observed after the second stage of encystation (pII). Low levels of cpn60 persisted at 20 min, 90 min, 1 day, and 3 days after excystation, increasing after 6 days.

Excystation is necessary to initiate infection. When cysts are ingested they are exposed to an increase in ambient temperature of greater than 30°C and a large decrease in pH as they pass from cold water into the host stomach (1). Cysts exposed to these stimuli during the induction stage of excystation (pI) *in vitro* displayed a constant level of cpn60 protein (Fig. 2*B*). A dramatic drop in the level of cpn60 protein was observed after the second stage of excystation (pII) *in vitro*, which is induced by exposure to trypsin at pH 8 (modeling the passage of cysts into the host small intestine). Reduced cpn60 expression persisted for at least 3 days after excystation but returned to previous levels after 6 days (Fig. 2*B*). Expression of the *G. lamblia* taglin control was constant throughout both encystation and excystation (not shown), indicating that the trypsin treatment did not directly cause the decrease in cpn60 protein.

Expression of *G. lamblia* cpn60 After Stress. Since cpn60 proteins often are regulated in response to stresses (36), we examined cpn60 protein abundance (via Western blot analysis) in *G. lamblia* cells exposed to heat shock and ethanol treatment for varying durations and intensity known to induce stress response in this organism (37). Transfer of cells to 40°C and 43°C did not alter the level of cpn60 protein, and neither did exposure of cells to 3% ethanol. Exposure of cultured cells to DTT can induce a stress response because of accumulation of misfolded proteins in the endoplasmic reticulum (38). However, *G. lamblia* cells exposed to DTT also showed no change in the level of cpn60 protein. Collectively, these results suggest that the synthesis of *G. lamblia* cpn60 is not responsive to generalized stresses. This is consistent with the absence of a heat shock protein in the 60-kDa range shown by metabolic labeling (37).

Localization of the *G. lamblia* cpn60 Protein. In most eukaryotes, a cpn60 protein is localized in mitochondria and facilitates refolding of proteins after transport across mitochondrial membranes (36). In trichomonads, cpn60, cpn10, and hsp70 localize to the hydrogenosome, an unusual double-membraned organelle of anaerobic energy metabolism (9). However, *G. lamblia* lacks recognizable mitochondria or hydrogenosomes. Preliminary evidence suggests that a subcellular compartment (possibly a mitochondrial relic) may exist in the amitochondriate parasite *Entamoeba histolytica* (reviewed in ref. 15). It is possible that the cpn60 protein in *G. lamblia* may be targeted to a similar structure.

Organelle-targeted proteins often have N-terminal extensions that are cleaved during transport into the organelle. An alignment of the *G. lamblia* and *E. histolytica* cpn60 N-terminal regions with other homologs is shown in Fig. 3. The processed targeting peptides of the mitochondrial and hydrogenosomal cpn60s of *Leishmania tarentolae* and *Trichomonas vaginalis* are 8 and 15 aa in length, respectively (9, 39). The mature N termini of these proteins correspond roughly with the N termini of GroEL homologs in eubacteria. *E. histolytica* cpn60 possesses a 8- to 9-residue, serine-rich extension relative to eubacterial homologs that resembles an N-terminal extension

```
G. lamblia          MLQHYTSVISGEDARSGLL...
E. histolytica      MLSSSSHYNGKLLSLNIDCRENVL...
L. tarentolae       MLRSAVRLAGKDVRFGEDARRSMQ...
T. vaginalis        MSLIEAAKHFTRAFAKARDLKFGSDARDHLL...
C. crescentus       MAAKDVYFSSDARDKML...
                          *  *****  **
```

FIG. 3. An alignment of the N termini of *G. lamblia*, *E. histolytica*, mitochondrial, hydrogenosomal, and eubacterial cpn60 homologs. The deduced N-terminal sequence of *G. lamblia* and *E. histolytica* cpn60s are aligned with homologs from *L. tarentolae*, *T. vaginalis*, and *Caulobacter crescentus*. The *L. tarentolae* and *T. vaginalis* targeting peptides (underlined) are removed during import into mitochondria and hydrogenosomes, respectively (9, 39). An N-terminal extension of the *E. histolytica* cpn60 homolog is suggestive of a targeting peptide for a cryptic organelle in this organism. *G. lamblia* cpn60 has a small, 2-aa N-terminal extension relative to *C. crescentus*. Amino acid identities of the *G. lamblia* cpn60 to other homologs are indicated by asterisks (*) under the alignment.

encoded by the pyridine nucleotide transhydrogenase gene of this organism (15). In contrast, the *G. lamblia* cpn60 extends only 2–3 aa past the N termini of eubacterial homologs and is probably too short to represent a full targeting signal. However, because similarity to other cpn60 and GroEL homologs does not start until amino acid position 8 of the *G. lamblia* cpn60 (Fig. 3), we cannot rule out the possibility that the first 7 aa, or a subset of them, constitute a targeting peptide to an unknown organelle in *G. lamblia*.

We employed immunoelectron microscopy to localize the cpn60 protein in *G. lamblia* cells. The *Synechococcus* sp. anti-cpn60 antibody yielded a punctate labeling pattern dispersed throughout the *G. lamblia* cytosol (not shown), in agreement with a previous report (19). Neither study detected an association of anti-cpn60 antibodies with any specific membranous compartment. Further studies are needed to better determine the localization and function of the cpn60 protein within *G. lamblia* cells.

The Phylogenetic Position of the *G. lamblia* cpn60 Homolog. An initial phylogenetic analysis was conducted on an alignment of 47 sequences of eubacterial, archaebacterial, eukaryotic cytosolic, and organellar cpn60 homologs. In the tree of highest log likelihood, the *G. lamblia* homolog shares a most recent common ancestor with mitochondrial-like cpn60 homologs from other eukaryotes (not shown). The relative branching order of the Gram-positive eubacterial, cyanobacterial, bacteroides, spirochete, chlamydial, proteobacterial, and mitochondrial lineages is similar to published small subunit rRNA phylogenies (40).

A second dataset was assembled containing cpn60 homologs that represent the diversity of mitochondrial and proteobacterial sequences, as well as their nearest outgroups. In the optimal trees recovered by protein distance, maximum parsimony, and protein maximum likelihood methods, the *G. lamblia* cpn60 homolog formed a clade with eukaryotic, mitochondrial-like cpn60s to the exclusion of the bacterial lineages (Fig. 4). In agreement with previous analyses (10, 12), all methods showed a specific relationship between the mitochondrial lineage and the subdivision of the α-Proteobacteria that contains the rickettsias.

Bootstrap support for the monophyly of the clade containing *G. lamblia* and mitochondrial-like sequences was strong when using protein maximum likelihood (100%) and protein distance (94%) methods, and weak (37%) when using the maximum parsimony method. The branching order within the

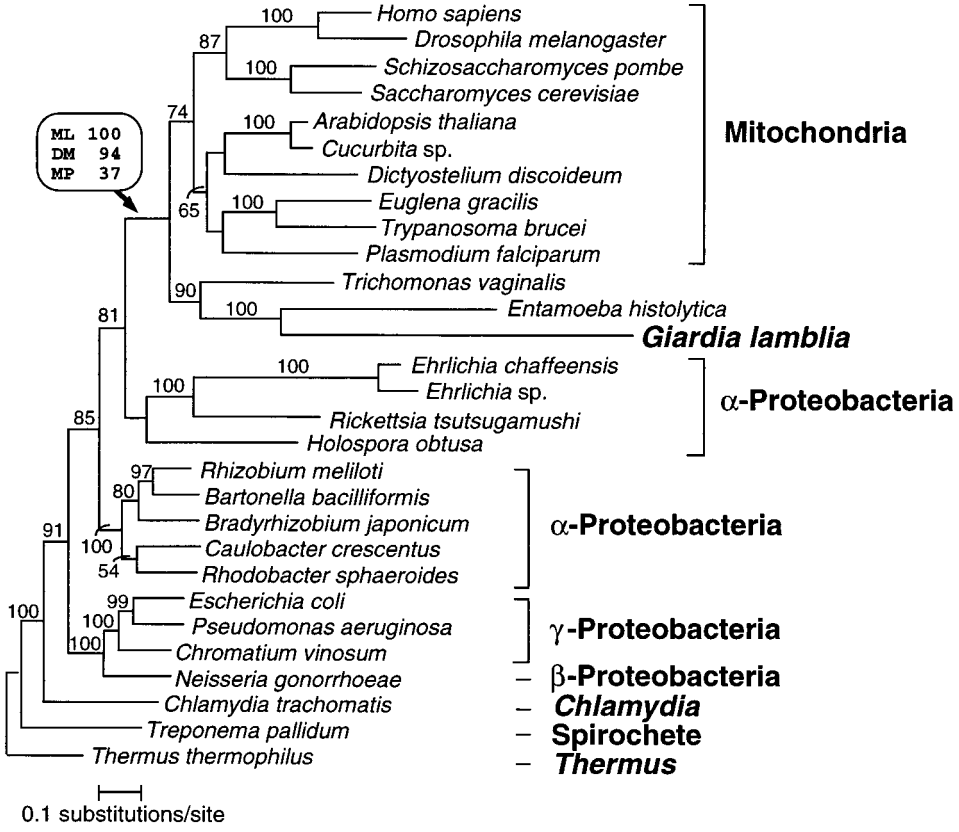

FIG. 4. Phylogenetic relationships of cpn60 homologs. Protein maximum likelihood analysis of 513 aligned amino acid positions yielded the tree shown (log likelihood = −18676.09). Optimal trees obtained by using protein distance and maximum parsimony methods differed from this topology in the branching order of the major eukaryotic groups and, to a lesser extent, within the α-, β-, and γ-proteobacterial clades. The branching order between these clades was identical for all methods. Bootstrap values obtained for the *G. lamblia*/mitochondrial-like cpn60 clade when using protein maximum likelihood (ML), protein distance matrix (DM), and maximum parsimony (MP) methods are shown in a box above the relevant node (indicated by an arrow). For all other nodes in the tree, protein ML bootstrap values (where >50%) are shown above each branch. The scale bar indicates estimated sequence divergence per unit branch length.

mitochondrial subtree varied according to the phylogenetic methods used and in many cases was not strongly supported by bootstrap analysis. However, several consistent features were apparent in protein maximum likelihood and distance trees, including the grouping of Metazoa with Fungi and the early divergence of the three amitochondriate groups (Fig. 4), in agreement with phylogenies of other molecules (2, 5, 41). Curiously, the cpn60s of the three amitochondriate protists, *G. lamblia*, *E. histolytica*, and *T. vaginalis*, form a clade in optimal trees obtained with each method. The *G. lamblia*/*E. histolytica* affinity seemed particularly strong with 100%, 92%, and 94% bootstrap support from protein maximum likelihood, distance, and maximum parsimony methods, respectively. This strong association was not expected because an affinity between *G. lamblia* and *E. histolytica* has not been observed in phylogenies of other genes such as small subunit rRNA, and elongation factors (2, 5). Because of this and the extremely divergent nature of both sequences, we suspect that this clade is an artifact. Although maximum likelihood methods are generally robust to the long branch attraction artifact, they may succumb to this problem when the evolutionary model is violated (42).

On further investigation, two violations of the maximum likelihood model became apparent. First, the amino acid frequencies in each sequence were compared with the overall frequencies in the dataset (incorporated in the model) by using a χ^2 test. Of the 29 sequences in the dataset, only the *G. lamblia*, *E. histolytica*, and *Thermus thermophilus* sequences deviated significantly from the overall amino acid frequencies in the dataset ($P < 0.0001$ for *G. lamblia*, $P < 0.02$ for *E. histolytica*, and $P < 0.02$ for *T. thermophilus*). For 11 of the 20 amino acid types, the *G. lamblia* and *E. histolytica* sequences deviated from the overall frequencies in the same direction.

A second violation of the protein maximum likelihood model involved rate variation among sites in the cpn60 dataset. The maximum likelihood estimate of the gamma shape parameter (α) for the cpn60 amino acid dataset was 0.69, suggesting that rate variation among sites in this dataset is extreme (43).

Under the hypothesis that these two model violations coupled with high rates of substitution in the *G. lamblia* and *E. histolytica* lineages were responsible for the artifactual clustering of these sequences, we explored the effect of corrections for these problems. To combat the effects of biased amino acid composition, we analyzed the nucleotide sequences coding for the cpn60 proteins. Second, we used distance and likelihood methods with a model (the HKY+Γ model) that accounts for rate variation among sites. Nucleotide distance analysis with trees selected under the minimum evolution criterion by using the HKY+Γ model (with Ti/Tv = 0.85 and $\alpha = 0.75$) showed that the *G. lamblia*/*E. histolytica* relationship was still recovered, but with lower bootstrap support (61%). By contrast, maximum-likelihood analysis by using the same model generated an optimal tree that did not display the *G. lamblia*/*E. histolytica* relationship. Bootstrap analysis by using this method on a taxonomically reduced dataset indicated that a *G. lamblia*/*E. histolytica* relationship is poorly supported (14% support). These results show that the *G. lamblia*/*E. histolytica* relationship observed in the amino acid phylogenies is probably artifactual. By contrast, optimal trees obtained with both nucleotide distance and maximum likelihood methods did recover a *G. lamblia*/*T. vaginalis* relationship, although it was poorly supported by bootstrap analysis (50% and 46% bootstrap support from distance and likelihood analyses, respectively). Support for the *G. lamblia*/mitochondrial-like cpn60 clade was somewhat stronger, gaining 66% bootstrap support from nucleotide distance analysis and 62% from nucleotide maximum likelihood analysis.

In addition to the nucleotide-level analyses, we reduced the effects of model violation and long-branch attraction artifacts by performing amino acid analyses on datasets with the divergent *E. histolytica* and *T. vaginalis* sequences removed. We observed strong bootstrap support for a *G. lamblia*/mitochondrial cpn60 clade in protein maximum likelihood (88%) and protein distance (82%) analyses, with lower support recovered by maximum parsimony methods (35%). Thus, the specific relationship of *G. lamblia* cpn60 to mitochondrial cpn60s is not an artifact because of the presence of the divergent *E. histolytica* and *T. vaginalis* cpn60 sequences.

The Evolutionary Origins of the *G. lamblia* cpn60 Gene. The presence of a cpn60 gene related to the mitochondrial cpn60 lineage in *G. lamblia* is most parsimoniously explained if it was transferred to the nucleus from mitochondria or their endosymbiotic ancestors. However, several other scenarios consistent with the phylogenetic data warrant consideration. First, it is possible that *G. lamblia* has acquired a cpn60 gene by lateral transfer from another eukaryotic lineage. This scenario could be distinguished from the previous one by establishing the presence or absence of cpn60 homologs in other diplomonads, such as the distantly related, free-living flagellate *Hexamita inflata* (2). If lateral transfer to the *Giardia* lineage occurred recently, then *Hexamita inflata* and its close relatives will lack this gene.

It is also possible that the ancestors of diplomonads acquired the cpn60 gene from an α-proteobacterial endosymbiont that was related to, but distinct from, the ancestors of mitochondria. This scenario could be tested if a free-living or endosymbiotic α-proteobacterium were found that contained a GroEL homolog that robustly grouped with the *G. lamblia* cpn60 in GroEL/cpn60 trees to the exclusion of the mitochondrial homologs.

In the absence of concrete evidence for either of these scenarios it is probable that the ancestors of *G. lamblia* acquired the cpn60 gene directly from the genome of the mitochondrial endosymbiont. The lack of mitochondrial functions in *G. lamblia* (44) is probably a result of secondary loss in early diplomonad evolution. The concomitant loss of many mitochondrion-targeted proteins may have caused a relaxation of functional constraints on diplomonad cpn60 proteins leading to their rapid divergence from other homologs. The long branches leading to both the *G. lamblia* and *E. histolytica* sequences in the cpn60 tree (Fig. 4) are therefore excellent examples of accelerated molecular evolution as a result of changed or reduced constraints on the function of a protein.

We cannot rule out the presence of a membrane-bounded mitochondrial relic organelle to which cpn60 and other proteins could be targeted. An organelle with a protein-import mechanism would rationalize the existence of a cpn60 homolog in *G. lamblia*, because cpn60 typically functions in refolding proteins during import into organelles in other eukaryotes (36). We do not know what the function of such an organelle would be in *G. lamblia*, but it might be related to energy metabolism. For instance, pyruvate:ferredoxin oxidoreductase (PFOR), an enzyme of anaerobic energy metabolism, functions within hydrogenosomes of trichomonads. In *G. lamblia*, PFOR is reported to be associated with membranes (45, 46), perhaps indicating the existence of a related organelle in this organism. A second possible function could be the detoxification of peroxide, because a membrane-associated NADH peroxidase exists in *G. lamblia* (47).

It is also possible that cpn60 may not localize within a membranous organelle but instead functions in the cytosol of *G. lamblia*. If other mitochondrion-derived proteins have been coopted for use in the *G. lamblia* cytosol, they may still require cpn60 to fold properly. Whatever its function, the expression of the cpn60 protein throughout much of the *G. lamblia* life cycle suggests that its presence may be essential to this organism.

CONCLUSIONS

The presence of a cpn60 gene of mitochondrial origin in *G. lamblia* suggests that diplomonads might not be representatives of a premitochondrial phase of eukaryotic evolution (8, 48). Instead, they could have lost mitochondrial functions secondarily. This conclusion is also supported by a recent finding that *G. lamblia*, along with other eukaryotes, possesses a proteobacterial-like valyl-tRNA synthetase that might derive from the mitochondrial endosymbiosis (T. Hashimoto, L. B. Sánchez, T. Shirakura, M. Müller, and M. Hasegawa, personal communication). Since diplomonads and trichomonads are among the earliest branching lineages in phylogenies of small subunit rRNA (2), elongation factors (5, 49), and the largest subunit of RNA polymerase II (4), these data suggest that the mitochondrial endosymbiosis may have occurred very early in eukaryote evolution. Two other flagellated protist groups remain as possible candidates for primitively amitochondrial eukaryotes: retortamonads and oxymonads (8, 48). If these groups are shown to branch with or later than diplomonads and trichomonads in molecular phylogenies, or they are shown to contain genes of mitochondrial origin, then the mitochondrial endosymbiosis may have taken place before the divergence of all known surviving eukaryotic lineages.

We thank H. Ward for antibodies to taglin, R. Gupta for antibodies to BiP, M. Hetsko for technical assistance, and J. M. McCaffery for ultrastructural analysis. We thank D. L. Swofford for allowing us to perform analyses with the PAUP* 4.0d56 program and publish the results. For critical reading of the manuscript, we thank A. G. B. Simpson. This work was supported by Grants AI24285, DK35108, and GM53835 awarded to F.D.G. from the National Institutes of Health, a grant from the Wellcome Trust awarded to C.G.C., Grant GM32964 awarded to M.L.S. from the National Institutes of Health, and the G. Unger Vettlesen Foundation. A.J.R. is supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada.

1. Adam, R. D. (1991) *Microbiol. Rev.* **55,** 706–732.
2. Leipe, D. D., Gunderson, J. H., Nerad, T. A. & Sogin, M. L. (1993) *Mol. Biochem. Parasitol.* **59,** 41–48.
3. Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & Peattie, D. A. (1989) *Science* **243,** 75–77.
4. Stiller, J. W. & Hall, B. D. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 4520- 4525.
5. Hashimoto, T. & Hasegawa, M. (1996) *Adv. Biophys.* **32,** 73–120.
6. Margulis, L. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1071–1076.
7. Cavalier-Smith, T. (1983) in *Endocytobiology II*, eds. Schwemmler, W. & Schenk, H. E. A. (De Gruyter, Berlin), pp. 1027–1034.
8. Patterson, D. J. (1994) in *Progress in Protozoology*, eds. Hausmann, K. & Hülsmann, N. (Gustav Fischer Verlag, Stuttgart), pp. 1–14.
9. Bui, E. T., Bradley, P. J. & Johnson, P. J. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9651–9656.
10. Horner, D. S., Hirt, R. P., Kilvington, S., Lloyd, D. & Embley, T. M. (1996) *Proc. R. Soc. Lond. B Biol. Sci.* **263,** 1053–1059.
11. Germot, A., Philippe, H. & Le Guyader, H. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14614–14617.
12. Roger, A. J., Clark, C. G. & Doolittle, W. F. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14618–14622.
13. Hirt, R. P., Healy, B., Vossbrinck, C. R., Canning, E. U. & Embley, T. M. (1997) *Curr. Biol.* **7,** 1–4.
14. Germot, A., Philippe, H. & Le Guyader, H. (1997) *Mol. Biochem. Parasitol.* **87,** 159–168.
15. Clark, C. G. & Roger, A. J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6518–6521.
16. Palmer, J. D. (1997) *Science* **275,** 790–791.
17. Henze, K., Badr, A., Wettern, M., Cerff, R. & Martin, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9122–9126.
18. Keeling, P. J. & Doolittle, W. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 1270–1275.
19. Soltys, B. J. & Gupta, R. S. (1994) *J. Parasitol.* **80,** 580–590.
20. Rosenthal, B., Mai, Z., Caplivski, D., Ghosh, S., de la Vega, H., Graf, T. & Samuelson, J. (1997) *J. Bacteriol.* **179,** 3736–3745.
21. Smith, M. W., Holmsen, A. L., Wei, Y. H., Peterson, M. & Evans, G. A. (1994) *Nat. Genet.* **7,** 40–47.
22. Que, X., Svärd, S. G., Meng, T. C., Hetsko, M. L., Aley, S. B. & Gillin, F. D. (1996) *Mol. Biochem. Parasitol.* **81,** 101–110.
23. Meng, T. C., Hetsko, M. L. & Gillin, F. D. (1996) *Infect. Immun.* **64,** 2151- 2157.
24. Hetsko, M. L., McCaffery, J. M., Svärd, S. G., Meng, T.-C., Que, X. & Gillin, F. D. (1998) *Exp. Parasitol.,* in press.
25. Ward, H. D., Lev, B. I., Kane, A. V., Keusch, G. T. & Pereira, M. E. (1987) *Biochemistry* **26,** 8669–8675.
26. Gupta, R. S., Aitken, K., Falah, M. & Singh, B. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2895–2899.
27. McCaffery, J. M., Faubert, G. M. & Gillin, F. D. (1994) *Exp. Parasitol.* **79,** 236–249.
28. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
29. Felsenstein, J. (1993) PHYLIP, Phylogeny Inference Package, Seattle (Univ. of Washington, Seattle), Version 3.57c.
30. Swofford, D. L. (1997) PAUP*, Phylogenetic Analysis Using Parsimony (*and other methods) (Sinauer, Sunderland, MA), Version 4.0d56.
31. Adachi, J. & Hasegawa, M. (1996) *Computer Science Monographs* (Institute of Statist. Math., Tokyo), No. 28.
32. Strimmer, K. & von Haeseler, A. (1996) *Mol. Biol. Evol.* **13,** 964–969.
33. Holberton, D. V. & Marshall, J. (1995) *Nucleic Acids Res.* **23,** 2945–2953.
34. Hilario, E. & Gogarten, J. P. (1995) *Biochim. Biophys. Acta* **128,** 94–98.
35. Das, A., Chiang, S., Fujioka, H., Zheng, H., Goldman, N., Aikawa, M. & Kumar, N. (1997) *Mol. Biochem. Parasitol.* **88,** 95–104.
36. Stuart, R. A., Cyr, D. M., Craig, E. A. & Neupert, W. (1994) *Trends Biochem. Sci.* **19,** 87–92.
37. Lindley, T. A., Chakraborty, P. R. & Edlind, T. D. (1988) *Mol. Biochem. Parasitol.* **28,** 135–144.
38. Pahl, H. K. & Baeuerle, P. A. (1997) *Trends Biochem. Sci.* **22,** 63–67.
39. Bringaud, F., Peyruchaud, S., Baltz, D., Giroud, D., Simpson, L. & Baltz, T. (1995) *Mol. Biochem. Parasitol.* **74,** 119–123.
40. Olsen, G. J., Woese, C. R. & Overbeek, R. (1994) *J. Bacteriol.* **176,** 1–6.
41. Wainright, P. O., Hinkle, G., Sogin, M. L. & Stickel, S. K. (1993) *Science* **260,** 340–342.
42. Lockhart, P. J., Larkum, A. W., Steel, M., Waddell, P. J. & Penny, D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1930–1934.
43. Yang, Z. (1996) *Trends Ecol. Evol.* **11,** 367–372.
44. Müller, M. (1988) *Annu. Rev. Microbiol.* **42,** 465–488.
45. Ellis, J. E., Williams, R., Cole, D., Cammack, R. & Lloyd, D. (1993) *FEBS Lett.* **325,** 196–200.
46. Townson, S. M., Upcroft, J. A. & Upcroft, P. (1996) *Mol. Biochem. Parasitol.* **79,** 183–193.
47. Brown, D. M., Upcroft, J. A. & Upcroft, P. (1995) *Mol. Biochem. Parasitol.* **72,** 47–56.
48. Cavalier-Smith, T. (1993) *Microbiol. Rev.* **57,** 953–994.
49. Yamamoto, A., Hashimoto, T., Asaga, E., Hasegawa, M. & Goto, N. (1997) *J. Mol. Evol.* **44,** 98–105.