Widespread and Ancient Distribution of a Noncanonical Genetic Code in Diplomonads

Patrick J. Keeling[1] *and W. Ford Doolittle*
Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia

Recently, a group of diplomonads has been found to use a genetic code in which TAA and TAG encode glutamine rather than termination. To survey the distribution of this characteristic in diplomonads, we sought to identify TAA and TAG codons at positions where glutamine is expected in genes for alpha-tubulin, elongation factor-1α, and the gamma subunit of eukaryotic translation initiation factor-2. These sequences show that the variant genetic code is utilized by almost all diplomonads, with the genus *Giardia* alone using the universal genetic code. Comparative phylogenetic analysis reveals that the switch to this genetic code took place very early in the evolution of diplomonads and was likely a single event. Termination signals and downstream untranslated regions were also cloned from three *Hexamita* genes. In all three of these genes, the predicted TGA termination codon was found at the expected position. Interestingly, the untranslated regions of these genes are high in AT. This is incongruent with the coding regions, which are comparatively GC-rich.

Introduction

Diplomonads are flagellated protozoa that lack identifiable mitochondria, peroxisomes, and permanent Golgi dictyostomes. These features are found in almost all eukaryotes, and their absence in diplomonads and a handful of other protists led to the proposal that these taxa must have diverged from other eukaryotes before any of these cellular structures arose (Cavalier-Smith 1983). For diplomonads, this view has garnered support from phylogenetic analyses of several molecules, which show them to be the deepest branching, or among the deepest branching, eukaryotes known (Cavalier-Smith 1993; Leipe et al. 1993; Hashimoto et al. 1994). However, whether or not these structures really were absent in the ancestor of diplomonads or have since been lost remains to be seen (see Keeling and Doolittle [1997] for discussion).

While the phylogenetic position of diplomonads within the eukaryotes has been studied in some depth with molecular data, phylogeny within the group has been addressed predominantly by morphological analyses. Molecular phylogenetic methods have been applied to the genus *Giardia,* the causative agent of an enteric disease in humans (van Keulen et al. 1993), but only recently has there been an effort to determine the relationships between other genera by molecular means (Branke et al. 1996; Cavalier-Smith and Chao 1996; Keeling and Doolittle 1996*a*; Rozario et al. 1996).

Indeed, most of the attention paid to the molecular biology of diplomonads has been concentrated on *Giardia,* but diplomonads are very diverse, including a wide variety of other pathogens and several free-living forms (for review see Kulda and Nohýnková 1978; Vickerman 1990). Recently, the first protein-coding genes from genera besides *Giardia* were characterized, and it was found that certain species do not use the "universal" genetic code (Keeling and Doolittle 1996*a*; Rozario et al. 1996). The genetic code used by these organisms differs from the universal genetic code in that TAA and TAG encode glutamine rather than signaling the termination of peptide synthesis. This has been shown most convincingly in the salmon parasite *Hexamita* 50330, where TAA and TAG codons have been found at highly conserved glutamine positions of several genes and genes for UAA- and UAG-decoding tRNAs have also been identified in the genome. Suggestive evidence was also found in the free-living diplomonads *Hexamita inflata* and *Trepomonas agilis*: UAA- and UAG-decoding tRNA genes were identified in the former (Keeling and Doolittle 1996*a*), and the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene in the latter contains a single in-frame TAA codon (Rozario et al. 1996).

Phylogenetic trees of diplomonads are also consistent with this distribution of genetic codes. In each case where they are represented, *T. agilis, H. inflata,* and *Hexamita* 50330 seem to comprise a lineage distinct from *Giardia,* which uses the universal genetic code (Cavalier-Smith and Chao 1996; Keeling and Doolittle 1996*a*; Rozario et al. 1996). However, the molecular data sets from which these trees were inferred do not share comparable representations of taxa. This makes an overall view of diplomonad phylogeny impossible without considering numerous trees at once, a danger when trees may not even be congruent. With a clear picture of diplomonad phylogeny, it should be more apparent when the alternative code arose and how prevalent it is.

Here we have sought to define the distribution of genetic codes within the diplomonads by three means. First, we have provided additional evidence that TAA and TAG encode glutamine by characterizing the carboxy terminus of several *Hexamita* genes. To date, the only termination codons that have been sequenced from any diplomonad are from *Giardia.* If TAA and TAG codons encode glutamine in these *Hexamita* species, then all legitimate termination codons in their genomes ought to be TGA. Second, by sequencing additional protein-coding genes from new species, the presence of

[1] Present address: Plant Cell Research Centre, School of Botany, University of Melbourne, Australia.

Key words: diplomonad, phylogeny, genetic code, rRNA, EF-1α, alpha-tubulin.

Address for correspondence and reprints: W. Ford Doolittle, Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia, Canada, B3H 4H7.

Mol. Biol. Evol. 14(9):895–901. 1997

TAA and TAG codons can be used as an indicator of which taxa use the noncanonical code. To this end, we have sequenced fragments of the gamma subunit of translation initiation factor-2 (eIF-2γ) from *H. inflata* and *Spironucleus vortens* and of the alpha-tubulin and elongatoin factor-1α (EF-1α) from *S. vortens*, and completed the previously truncated EF-1α from *S. muris*. Finally, we have sequenced the small-subunit ribosomal RNA (rRNA) genes from *Hexamita* 50330 and *S. vortens*. These genes, along with the protein-coding genes reported here, help to "even-up" the taxon sampling of alpha-tubulin, EF-1α, and small-subunit ribosomal RNA so that the trees inferred from these molecules can be compared.

Together, these new data reveal that this alternative genetic code is used by *all* diplomonads for which molecular data are presently known, with the single exception of *Giardia*. Moreover, comparing trees of small-subunit rRNA, EF-1α, and alpha-tubulin shows that *Giardia* is the earliest branch of diplomonads and that the noncanonical genetic code arose just once, early in the evolution of diplomonads.

Materials and Methods

Strains, Culture Conditions, and Molecular Techniques

Hexamita strains ATCC 50330 and 50380 were grown as described previously (Keeling and Doolittle 1996*a*). There is some discrepancy in the literature concerning the names of these two diplomonads. Alpha-tubulin genes have been sequenced from both, and these genes differ by only three synonymous substitutions (Keeling and Doolittle 1996*a*), suggesting that the two strains are likely variants of the same species. To distinguish these strains here, they will be referred to as *Hexamita* 50330 and *Hexamita* 50380. This is consistent with ATCC and our previous work on the genetic code (Keeling and Doolittle 1996*a*), although they have also been called *Spironucleus* sp. by Rozario et al. (1996), and sequences in GenBank are cataloged as Diplomonad ATCC50330 and Diplomonad ATCC50380. Phylogenetic data presented in this work will show that they likely belong to neither the genus *Hexamita* nor the genus *Spironucleus* and most likely should be given the distinction of a unique genus when complete morphological descriptions have been made.

Spironucleus vortens (ATCC 50386) was cultured axenically in Keister's Modified TYI-S-33 medium at 30°C in the dark. This medium is much more easily made than the *S. vortens* medium originally discribed (Poynton et al. 1995), but we found that the growth rate and maximum cell density reached were quite satisfactory, far exceeding that of either *Hexamita* 50330 or *Hexamita* 50380. DNA was isolated from these organisms by repeated extractions of lysed cells with phenol chloroform and cetyltrimethylammonium bromide (CTAB). DNA from *H. inflata* AZ-4 and *S. muris* were provided by H. van Keulen.

All molecular manipulations were carried out using *Escherichia coli* strain DH5αF′, which was grown in LB broth and on LB agar with ampicillin selection. Amplification products were purified by gel isolation, cloned using the TA vector pRC2.2 (Invitrogen), and sequenced either manually or using LiCor or ABI automated sequencing protocols. Full-length sequence was obtained by sequencing fragments subcloned into p-Bluescript SK+ (Stratagene) or by gap-filling with primers.

Amplifying Genes for Alpha-Tubulin, EF-1α, and Small-Subunit rRNA

Alpha-tubulin genes were amplified using primers TCCGAATTCARGTNGGAAYGCNTGYTGGGA and TCCAAGCTTCCATNCCYTCNCCNACRTACCA (provided by A. J. Roger), small-subunit rRNA using primers CUACUACUACUACAACCTGGTTGATCCTGCCAGT and CUACUACUACUAGATCCTTCTGCAGGTTCACCTAC (provided by M. Ragan), eIF-2γ using primers CGCCAGGCCACSATHAAYATHGGNAC and ATCAGGATGTCRTGNCCNGGRCARTC for *H. inflata* and primers CGCCAGGCCACSATHAAYATHGGNAC and CCGCCTGGCTTGTTNACRTCRAA for *S. muris*, and EF-1α using primers CAACATCGTCGTCATCGGNCAYGTNGA and GCCGCGCACGTTGAANCCNACRTTRTC. The C-terminal portion of the *S. muris* EF-1α was amplified using primers TGATGCCATCGACGGACTCAAGGC and GCCGCGCACGTTGAANCCNACRTTRTC. Alpha-tubulin and EF-1α amplifications consisted of 10 cycles with an annealing temperature of 35°C and 25 cycles with an annealing temperature of 45°C. eIF-2γ amplifications consisted of 30 cycles with an annealing temperature of 57°C. Small-subunit rRNA amplifications consisted of 30 cycles with an annealing temperature of 50°C. All reactions included *Taq* DNA polymerase as well as *Pfu* DNA polymerase to correct errors. Two to four clones of each amplification product were sequenced entirely, in the case of the 3′ end of *H. inflata* EF-1α from three separate reactions, to detect ambiguities, and none were found except in the small-subunit rRNA genes, which appear to have amplified from different copies of the gene in both *Hexamita* 50330 and *S. vortens*.

Single-Primer Amplifications

Clones covering the termination codons of the alpha-tubulin genes from *H. inflata* and *Hexamita* 50380 and of the EF-1α gene of *H. inflata* were obtained by amplification from genomic DNA using a single primer. This primer was designed to match the coding region of the gene in question, and reactions were carried out assuming that in the pool of single-primer product that would result, those products corresponding to the target gene would be highly represented. Reactions consisted of 35 cycles with an annealing temperature of 35°C to promote nonstringent binding. The products of these reactions were then separated by agarose electrophoresis, and for each reaction a single abundant fragment calculated to be large enough to extend beyond the 3′ terminus of the gene (according to the position of the primer used and the lengths of homologes from other organisms) was isolated, cloned, and sequenced. In each of the three instances, this fragment proved to correspond to the 3′ end of the expected gene. This strategy was used to obtain the 3′ end of EF-1α from *H. inflata* using the primer GACAAGCCACTCCGTCTCCCA, of alpha-tubulin from *H. inflata* using the primer AGCCCGCAAACAT-GATGGTCAAG, and of alpha-tubulin from *Hexamita* 50380 using the primer CCCAGATGATCTCTGGTAT-GACTGC.

Phylogenetic Analysis

Alpha-tubulin and EF-1α alignments were constructed using the PileUp program from the GCG package (Devereux, Haeberli, and Smithies 1984), and the alignments were edited by eye. A subset of the small-subunit rRNA alignment from the Ribosomal Database Project (Maidak et al. 1996) was downloaded, and additional diplomonad rRNAs were aligned to this template. Phylogenetic trees of these three molecules were inferred using parsimony, distance, and maximum likelihood. Alignments and some trees are not shown due to their large number, but both can be obtained by request from the authors.

In the case of the protein-coding genes, amino acid sequences were used for unweighted parsimony (conducted using PAUP version 3.1.1; Swofford 1993) and for corrected distances, which were calculated according to the PAM 250 substitution matrix and used to construct neighbor-joining trees (using PROTDIST and NEIGHBOR programs from the PHYLIP 3.5c package; Felsenstein 1993). Bootstrap support for all trees was calculated by conducting 100 replicates with resampling. An EF-1α data set consisting of 40 taxa and 420 positions was used which included archaebacterial outgroups. For alpha-tubulin, 36 taxa and 406 positions were used with no outgroup, since other tubulin paralogues are quite distant and rooting with them is dubious (see Keeling and Doolittle 1996*b*).

Protein maximum-likelihood trees were inferred by constraining groups of organisms of known relationship. For EF-1α, 420 positions were used with 19 taxa divided into seven groups: animals and fungi, plants, euglenozoa, *Dictyostelium, Entamoeba,* archaebacteria, and diplomonads. For alpha-tubulin the data consisted of 384 positions and 18 taxa in eight groups: animals, plants, fungi, microsporidia, parabasalia, heterolobosea, euglenozoa, and diplomonads. These groups were exhaustively searched using the PROTML program (from the MOLPHY 2.2 package; Adachi and Hasegawa 1992) with likelihood calculations and statistical tests based on the JTT-F transition probability matrix.

Phylogenetic trees based on small-subunit rRNA genes were also inferred using parsimony (again with PAUP version 3.1.1), distance (using DNADIST and NEIGHBOR), and maximum likelihood using the DNAML program (all from the PHYLIP 3.5c package). In parsimony and neighbor-joining analyses, the data consisted of 36 taxa and 1,036 positions, and in maximum-likelihood analysis, the same positions of 28 taxa were used. Of the 12 diplomonad small-subunit genes available to us, *Hexamita* sp. (GenBank accession number Z17224), *Giardia intestinalis* (GenBank accession number X52949), *S. vortens*-2, and *Hexamita* 50330-2 were not used, as they differ by only 13, 2, 5, and 8 substitutions, respectively, from other genes represented in the tree. Trees consisting of only the diplomonads were also inferred for all three molecules by all three methods, to see if the position of the root within the phylum was otherwise affecting the topology, but in no case was a significant difference observed.

Results and Discussion

Termination Codon Use in *Hexamita* Genes

In genomes where TAA and TAG code for glutamine, the bona fide termination signal of all protein-coding genes should be TGA. However, among the diplomonads, full-length genes including the termination codon have only been sequenced from *Giardia.* Collectively, these *Giardia* genes use all three canonical termination codons in a more or less random proportion (see the TransTerm Database, which currently has 45 *Giardia* genes; Dalphin et al. 1996), and in no case has either a TAA or TAG codon been found within a coding sequence. One *Giardia* gene has been argued to contain a TAG codon that is translated (Upcroft et al. 1990), but this is more likely a legitimate termination codon, as the evidence for its translation is based on read-through in *Escherichia coli* and not *Giardia* itself.

To identify the actual termination signals from organisms that use the noncanonical code, single-primer PCR was used to isolate the 3′ ends of the alpha-tubulin genes from *Hexamita* 50380 and *H. inflata,* and the EF-1α from *H. inflata.* In each case, the PCR product had a considerable overlap with the previously identified gene, and in both alpha-tubulins, this overlap was identical (over 523 and 328 bp for *Hexamita* 50380 and *H. inflata,* respectively). The 556-bp overlap between the single-primer PCR product of *H. inflata* EF-1α and the previously reported EF-1α gene contained two mismatches, one synonymous and the other resulting in a tryptophan-to-cysteine substitution. Since the cysteine residue is highly conserved at this position of EF-1α genes, it is likely that the original clone has an error here.

Following this overlap, each of these clones also contained a length of new coding sequence, followed by the predicted TGA termination codon. No TAA or TAG codons appear in the coding sequence immediately upstream of the TGA codon, and each TGA codon appears at a position close to the ends of homologous genes from other organisms (fig. 1), leaving little doubt that these are the actual termination codons of these genes.

The surprising feature of all these sequences is the relatively high AT content of the noncoding sequence that follows the termination codon in genomes that are thought to be rather GC-rich. The proportions of G and C residues at all synonymous positions where GC can vary (synonymous third positions plus the first position of arginine and leucine codons) in these particular genes are 55% and 65% for *Hexamita* 50380 and *H. inflata,* respectively, and yet the noncoding regions immediately downstream are only 31% and 33% GC, respectively. It may be that the elevated AT content of these regions is

A. EF-1α

Hexamita inflata	DMKRTVAVGVVTEVLKKDK•
S. lemnae	DMKQTVAVGVIKEVVKKEQKGMVTKAAQKKK•
T. pyriformis	DMKQTVAVGVIKKVEKKDK•
E. crassus 1	DMRQTVAVGVIQEIKKKATEDKKGKKK•
E. crassus 2	DMKRTVAVGVIQEVIHKKETKKKASKR•
Homo	DMRQTVAVGVIKAVDKKAAGAGKVTKSAQKAQKAK•
Arabidopsis	DMRQTVAVGVIKSVDKKDPTGAKVTKAAVKKGAK•
Dictyostelium	DMRQTVAVGVIKSTVKKAPGKAGDKKGAAAPS•
Plasmodium	DMRQTIAVGIINQLKRKNLGAVTAKAPAKK•
Trypanosoma	DMRQTVAVGIIKAVTKKDGSGGKVTKAAVKASKK•
Saccharomyces	DMRQTVAVGVIKSVDKTEKAAKVTKAAQKAAKK•
Sulfolobus	DMGKTVGVGVIIDVKPRKVEVK•
Methanococcus	DMGMTVAAGMAIQVTAKNK•

B. Alpha-Tubulin

Hexamita inflata	WYVSEGMEEGEFAEAREDLAALEKDYEEIGADTVAQGEGEGEDM•
Hexamita 50380	WYVGEGMEEGEFSEAREDLASLEKDYEEIGQDTVADGEGEGGEEDY•
O. granulifera	WYVGEGMEEGEFSEAREDLAALEKDYEEVGIETAEGEGEEEGME•
T. pyriformis	WYVGEGMEEGEFSEAREDLAALEKDYEEVGIETAEGEGEEEGY•
T. thermophila	WYVGEGMEEGEFSEAREDLAALEKDYEEVGIETAEGEGEEEGY•
S. lemnae 2	WYVGEGMEEGEFSEAREDLAALEKDYEEVGIETAEGEGEEEGME•
S. lemnae D	WYVGEGMEEGEFSEVREDLAALEKDYEEVGIEIVEGEGEEEGME•
E.vannus	WYVGEGMEEGEFSEAREDLAALEKDYEEVGIETAEGEGEEEDMA•
E. octocarinatus	WYVGEGMEEGEFSEAREDLAALEKDYEEVGIETAEGEGEEEGME•
Homo	WYVGEGMEEGEFSEAREDMAALEKDYEEVGVHSVEGEGEEEGEEY•
Arabidopsis 1	WYVGEGMEEGEFSEAREDLAALEKDYEEVGGEGAEDDDEEGDEY•
Plasmodium	WYVGEGMEEGEFSEAREDLAALEKDYEEVGIESNEAEGEDEGYEADY•
Trypanosoma	WYVGEGMEEGEFSEAREDLAALEKDYEEVGAESADMDGEEDVEEY•
Saccharomyces	WYVGEGMEEGEFTEAREDLAALERDYIEVGADSYAEEEEF•
Naegleria	WYVGEGMEEGEFSEAREDLAALEKDYEEVGTESQEGDGEEGEDGGDQ•
Entamoeba	HYVGEGMEENEFTDARQDLYELEVDYANLALDNTIEGESMTAQ•

FIG. 1.—Alignment of inferred carboxy-termini of (*A*) EF-1α and (*B*) alpha-tubulin, with termination codons represented by a dot. Sequences from diplomonads (*Hexamita* 50380 and *H. inflata*) are aligned to various homologs, including numerous ciliates that also use genetic codes where termination codons have been affected (in *Euplotes*, TGA encodes cysteine, and in the other ciliates shown, TAA and TAG encode glutamine). The sequences shown begin at the positions from which the diplomonad sequences extend beyond the previously known sequences of these genes.

A. Dynein Beta Chain

		238		239		260	
Tripneustes	QRWPLMIDP	Q	L	Q	GIKWIKQKYGDD-LRVIRIG	Q	RGYLDTIENAI
Anthocidaris	QRWPLMIDP	Q	L	Q	GIKWIKQKYGDE-LRVIRIG	Q	RGYLDTIENAI
Chlamydomonas	SRWALMIDP	Q	L	Q	GIKWIINKETNNGLVIIQQS	Q	PKYIDQVINCI
Paramecium	SRWPLIIDP	Q	L	Q	GSVWIRGSQGDN-LITINIS	Q	NKWLQQLNQAI
Hexamita inflata	QRWPLIIDP	*	L	*	GMTWIKKKEGSN-LKIVRFN	*	QGWMKEVERAL
		taa		**taa**		**taa**	

B. eIF-2γ

		129	
Saccharomyces	AGNESCP	Q	PQTSEHLAAI
S. pombe	AGNESCP	Q	PQTSEHLAAI
Drosophila	AGNESCP	Q	PQTSEHLAAI
Homo	AGNESCP	Q	PQTSEHLAAI
S.vortens	SAEQRCP	*	EQTREHFQAI
		tag	

C. EF-1α

		118	
S. vortens	ACEFTKFL	Q	KLNSRTLKP
Hexamita 50330	ACKFDAFL	Q	KLNARTLKP
Giardia	ACQFQLFL	Q	KLDKRTLKP
S.muris	ACKFEKFL	*	KIDQRTMKP
		tag	

FIG. 2.—Inferred amino acid sequences surrounding TAA and TAG codons in (*A*) *H. inflata* dynein beta-heavy chain, (*B*) *S. vortens* eIF-2γ, and (*C*) *S. muris* EF-1α. Positions marked by a Q represent canonical CAA or CAG codons, while positions marked by an asterisk represent TAA or TAG codons. Numbering refers to the codon number in the actual fragment in which the codons were found.

a reflection of AT-rich control elements such as transcription terminators or polyadenylation sites, but the regions appear to be uniformly AT-rich over several hundred base pairs.

The incongruency between coding and noncoding DNA should be considered with caution, however, as there are scarcely enough data to make inferences about the mutation pressures on the genome or even on the codon biases. There are only a handful of genes from any of these genomes, and the synonymous GC proportions of even the genes that are known are extremely variable between organisms and even between two genes from one organism. For instance, all currently known *Hexamita* 50330 genes use G or C in 70% of synonymous positions, while the synonymous GC proportion of the single gene known from *T. agilis* is only 19%. Similarly, the synonymous GC proportion of the *H. inflata* dynein gene is only 24%, but the *H. inflata* alpha-tubulin, EF-1α, eIF-1γ, and GAPDH genes are all GC-rich. It would be overly simplistic, and perhaps misleading, to attempt to distill these observations down to the pressures on these genomes. At present, it may be sufficient to say that in the diplomonads where TAA and TAG encode glutamine, the canonical CAA and CAG codons seem to be favored over the "new" glutamine codons.

Protein-Coding Genes from Diplomonads Reveal the Widespread Use of the Noncanonical Genetic Code

Several lines of evidence indicate that in *Hexamita* 50330, TAA and TAG encode glutamine. Conversely, the large number of "normal" genes from *Giardia* is strong evidence that this genus uses the universal code. A number of new protein-coding genes were isolated and sequenced from *H. inflata, S. muris,* and *S. vortens* to identify any TAA and TAG codons that might show which additional taxa use this variant genetic code.

Hexamita inflata was previously shown to contain UAA- and UAG-decoding tRNAs, but neither TAA nor TAG codons in any of the protein-coding genes described (alpha-tubulin, EF-1α, or GAPDH). The reasoning previously applied to the absence of these codons is that they only occur at a very low frequency: highly expressed genes such as these tend to reflect a strong codon bias, so *H. inflata* genes that are likely less highly expressed ought to contain TAA or TAG codons. To investigate this possibility, a fragment of the gene for eIF-2γ was sequenced and was indeed found to contain a TAA codon, but at a position that is extremely variable in other homologes. However, more convincing support also comes from a recently identified fragment of the *H. inflata* dynein beta chain (GenBank accession number U82547; unpublished data) which contains 13 TAA and 2 TAG codons, many of which appear at positions otherwise highly conserved for glutamine (fig. 2*A*).

Spironucleus vortens is a recently described parasite isolated from the lips of the freshwater angelfish, *Pterophyllum scalare* (Poynton et al. 1995). No molecular data are known from *S. vortens,* so we amplified and sequenced the genes for EF-1α and alpha-tubulin. No termination codons were found in these genes, but once again when eIF-2γ was sequenced it was found to contain a TAG codon at a conserved glutamine position (fig. 2*B*).

Spironucleus muris has previously yielded neither termination codons in reading frames nor tRNAs that could decode them. Furthermore, the phylogenetic position of *S. muris* inferred by EF-1α argued that it would most likely use the universal code (Keeling and Doolittle 1996*a*). However, the EF-1α fragment used in these analyses covered only two thirds of the length of the gene, so we amplified a fragment that appears to be contiguous with the previously reported gene (they are identical over 537 bp of overlap). Surprisingly, we found that this fragment of EF-1α contains a TAG codon at a position conserved for glutamine in other diplomonads (fig. 2*C*).

Phylogenetic Distribution of Genetic Codes in the Diplomonads

The presence of this TAG codon in the EF-1α gene of *S. muris* is not easily reconciled with the phylogeny of EF-1α genes previously reported (Keeling and Doolittle 1996*a*). To determine if the use of this code is restricted to a single lineage or has evolved more than once as in ciliates (Baroin-Tourancheau et al. 1995), we compared EF-1α trees with those of other molecules. Presently there are only four molecules with more than two diplomonad genera represented: EF-1α, small-subunit rRNA, alpha-tubulin, and GAPDH; but the taxa represented in each of these data sets are not the same, so we have attempted to make the data more easily comparable by adding key taxa to each set. The branching position of the diplomonads within the eukaryotes has been addressed frequently and recently with all of these molecules (Leipe et al. 1993; van Keulen et al. 1993; Hashimoto et al. 1994; Branke et al. 1996; Cavalier-Smith and Chao 1996; Keeling and Doolittle 1996*b*; Rozario et al. 1996), so it will not be discussed here.

In addition to the new EF-1α and alpha-tubulin genes described above, we amplified and sequenced the small-subunit rRNA genes from *Hexamita* 50330 and *S. vortens*. Products of the expected size were sequenced in duplicate from each of these species, and in both cases the two sequences varied at a handful of positions. In the case of *Hexamita* 50330, the two genes differ at 8 out of 1,446 sites, and in the case of *S. vortens*, they differ at 5 out of 1,466 sites, all at variable positions in the small-subunit alignment. It is likely that each of these represents a separate copy of the gene, but the level of divergence was low enough not to warrant sequencing further copies.

Trees of small-subunit rRNA, alpha-tubulin, and EF-1α were inferred using maximum likelihood, neighbor-joining, and parsimony for diplomonads alone and also for diplomonads rooted by outgroups. Maximum-likelihood trees of each of these molecules are shown in figure 3. In each case, the topology of the diplomonads inferred by parsimony and distance matched that shown except for EF-1α, where the neighbor-joining tree varied in the relative order of *S. vortens, G. lamblia,* and *Hexamita* 50330, and in the parsimony trees of small-subunit rRNA and alpha-tubulin, which are not completely resolved. A consensus of these analyses suggests that the maximum-likelihood topologies shown in figure 3 are reasonable estimates for each molecule.

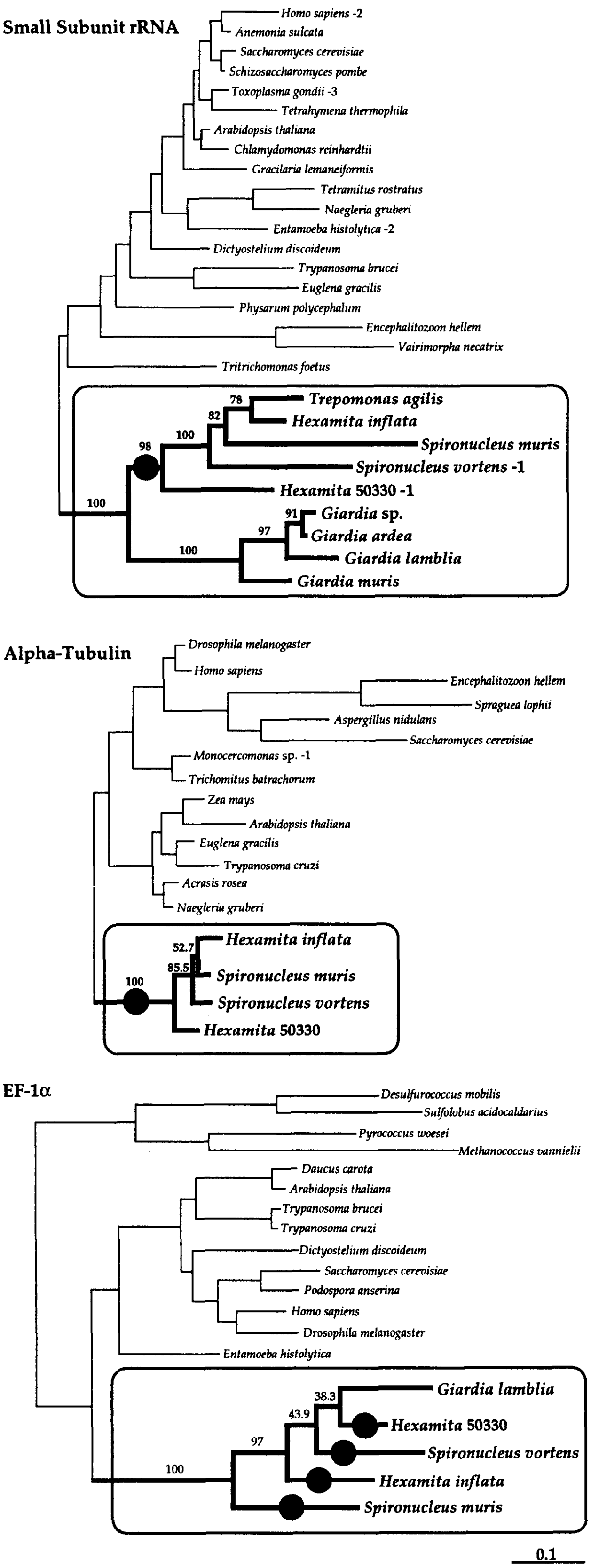

FIG. 3.—Maximum-likelihood trees of small-subunit rRNA, alpha-tubulin, and EF-1α. Scale represents 0.1 changes per site, and nodes at which changes to the genetic code must have occurred (according to that topology) are marked with a filled circle. In both small-subunit rRNA and alpha-tubulin a single evolution of the noncanonical code is required, while the topology inferred from EF-1α would call for four independent events if altering the code is considered irreversible. Alternatively, *Giardia* may have reverted to the universal code, but the most reasonable explanation remains that the root of the diplomonads in the EF-1α is incorrect.

←

The alpha-tubulin and small-subunit rRNA trees are congruent with one another but not with EF-1α (the GAPDH tree topology is also congruent with small-subunit rRNA and alpha-tubulin but not with EF-1α; Rozario et al. 1996). It is interesting, however, that the topology of the EF-1α tree is actually the same as the others, but the position of the diplomonad root is in the branch leading to *S. muris* rather than in that leading to *Giardia.* Since the *S. muris* EF-1α is relatively divergent, it is possible that the position of the root in this tree is erroneous. The positions of the roots in the branches leading to *Giardia* or *S. muris* were compared to one another directly by maximum likelihood in both small-subunit rRNA and EF-1α. These tests show that neither root is significantly superior to the other for EF-1α data (the two alternatives being within 1.55 SE), but in the case of small-subunit rRNA, rooting the diplomonads in the branch leading to *Giardia* is favored over rooting them in the branch leading to *S. muris* by 4.66 SE. Rooting the diplomonads in the branch leading to *Giardia* also results in the most parsimonious distribution of genetic codes (fig. 3), so, altogether, this topology does appear to be the more likely.

Implications for the Evolution of Diplomonads and the Genetic Code

Traditional phylogenetic treatments of diplomonads based on ultrastructural characteristics predict the same tree topology as the small-subunit rRNA trees, but root the diplomonads differently based on the increasing degree of adaptation to parasitism observed in a series from free-living members such as *T. agilis* and *H. inflata* to specialized pathogens such as *Giardia* (Brugerolle 1975; Brugerolle and Taylor 1977; Siddall, Hong, and Dresser 1992). By contrast, molecular phylogeny and the distribution of genetic codes both argue that the presently free-living diplomonads diverged most recently (fig. 3). It is doubtful that a relatively versatile free-living phagotroph like *Trepomonas* could have developed from a highly adapted parasitic form like *Giardia* or *Spironucleus,* where cytostomes and other complex feeding apparatus were absent or severely reduced. It is perhaps more likely that individual parasitic types have adapted to parasitism independently. Molecular data from the flagellates thought to be closely related to the diplomonads, the retortamonads and oxymonads, should be sought to give a clearer impression of the ancestral state of this group (see Brugerolle 1993), and the presence or absence of the universal genetic code in these organisms will also be a strong character to complement phylogeny.

Although it seems that the alternative code appeared only once in the evolution of diplomonads, it is also clear that this same genetic code has evolved many times in different eukaryotic lineages. Of all changes to the genetic code in the nucleus, there is a single case where CTG encodes serine in certain species of *Candida* (Kawaguchi et al. 1989; Ohama et al. 1993) and a single case of TGA encoding cystein in the ciliate genus *Euplotes* (Meyer et al. 1991), but there are many cases of TAA and TAG encoding glutamine: in diplomonads, in the dasycladacean green algae *Acetabularia* and *Batophora* (Schneider, Leible, and Yang 1989; Schneider and de Groot 1991), and again in ciliates, where it is thought to have evolved more than once independently (Baroin-Tourancheau et al. 1995).

Various hypotheses have been put forward to explain the appearance of this code (Osawa et al. 1992; Schultz and Yarus 1994), and its prevalence in eukaryotes (Cohen and Adoutte 1995; Cedegren and Miramontes 1996). We have proposed that the frequent conversion of TAA and TAG from termination signals to glutamine codons may simply be a result of the fact that TAA and TAG and the canonical CAA and CAG glutamine codons require only a single transition in order to be interconverted (Keeling and Doolittle 1996*a*). The high frequency of these mutations would favor glutamine codons mutating to TAA and TAG, and such mutations are required by most of the models for changes to the genetic code to create an altered genetic code or, at the very least, to fix it in the population. This suggestion, however, does not diminish the possibility that other forces are also involved in alterations to the genetic code. Indeed, one of the most solid conclusions that we can make is the likelihood that the forces and mechanisms involved in altering the genetic code have been different in different genomes. Many changes have taken place independent of one another, and it is perhaps naive to assume that the same mechanism should be responsible for every instance of a phenomenon that is almost certainly simply the chance result of opportunity.

Acknowledgments

This work was funded by an operating grant (MT4467) from the Medical Research Council of Canada (MRC). P.J.K. was supported by a studentship from the MRC and, subsequently, a fellowship from the Natural Sciences and Engineering Research Council of Canada. W.F.D. is a fellow of the Canadian Institute for Advanced Research. We would like to thank H. van Keulen for providing DNA samples, A. J. Roger and M. A. Ragan for providing primers, and T. Cavalier-Smith for allowing us to use an unpublished *T. agilis* rRNA sequence. We would like to thank C. Chow for her assistance with sequencing, N. M. Fast for coordinating work between Australia and Canada, and M. Müller and G. I. McFadden for their interest and for comments on the manuscript. P.J.K. would also like to thank G. I. McFadden for allowing these analyses to take place in his laboratory. New sequences have been deposited in GenBank under the accession numbers U93083–U93089 and U94404–U94408.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1992. MOLPHY: programs for molecular phylogenetics, I. PROTML: maximum likelihood inference of protein phylogeny. Version 2.2. Computer Science Monographs, no. 27. Institute of Stastical Mathematics, Tokyo.

BAROIN-TOURANCHEAU, A., N. TSAO, L. A. KLOBUTCHER, R. E. PEARLMAN, and A. ADOUTTE. 1995. Genetic code deviations in the ciliates: evidence for multiple and independent events. EMBO J. **14**:3262–3267.

Branke, J., M. Berchtold, A. Breunig, H. König, and J. Reimann. 1996. 16S-like rRNA sequence and phylogenetic position of the diplomonad *Spironucleus muris* (Lavier 1936). Eur. J. Protistol. **32**:227–233.

Brugerolle, G. 1975. Contribution à l'étude cytologique et phylétique des diplozoaires (Zoomastigophorea. Diplozoa, Dangeard, 1910). VI. Charactères généraux des diplozoaires. Protozoologica **11**:111–118.

———. 1993. Evolution and diversity of amitochondrial zooflagellates. J. Eukaryot. Microbiol. **40**:616–618.

Brugerolle, G., and F. J. R. Taylor. 1977. Taxonomy, cytology and evolution of the Mastigophora. Pp. 14–28 *in* S. H. Hutner, ed. Proceedings of the Fifth International Congress of Protozoology. Pace University, New York, N.Y.

Cavalier-Smith, T. 1983. A 6-kingdom classification and a unified phylogeny. Pp. 1027–1034 *in* W. Schwemmler and H. E. A. Schenk, eds. Endocytobiology II: intracellular space as oligogenetic. de Gruyter, Berlin.

———. 1993. The kingdom Protozoa and its 18 phyla. Microbiol. Rev. **57**:953–994.

Cavalier-Smith, T., and E. E. Chao. 1996. Molecular phylogeny of the free-living archezoan *Trepomonas agilis* and the nature of the first eukaryote. J. Mol. Evol. **43**:551–562.

Cedegren, R., and P. Miramontes. 1996. The puzzling origin of the genetic code. Trends Biochem. Sci. **21**:199–200.

Cohen, J., and A. Adoutte. 1995. Why does the genetic code deviate so easily in ciliates? Biol. Cell **85**:105–108.

Dalphin, M. E., C. M. Brown, P. A. Stockwell, and W. P. Tate. 1996. TransTerm: a database of translational signals. Nucleic Acids Res. **24**:216–218.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. University of Washington, Seattle.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto, and M. Hasegawa. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. Mol. Biol. Evol. **11**:65–71.

Kawaguchi, Y., H. Honda, J. Taniguchi-Morimura, and S. Iwasaki. 1989. The codon CUG is read as serine in an asporogenic yeast *Candida cylindracea*. Nature **341**:164–166.

Keeling, P. J., and W. F. Doolittle. 1996*a*. A non-canonical genetic code in an early diverging eukaryotic lineage. EMBO J. **15**:2285–2290.

———. 1996*b*. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol. Biol. Evol. **13**:318–326.

———. 1997. Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. Proc. Natl. Acad. Sci. USA **94**:1270–1275.

Kulda, J., and E. Nohýnková. 1978. Flagellates of the human intestine and of other species. Pp. 1–138 *in* J. P. Kreier, ed. Parasitic protozoa. Vol. 2. Academic Press, New York, N.Y.

Leipe, D. D., J. H. Gunderson, T. A. Nerad, and M. L. Sogin. 1993. Small subunit ribosomal RNA of *Hexamita inflata* and the quest for the first branch of the eukaryotic tree. Mol. Biochem. Parasitol. **59**:41–48.

Maidak, B. L., G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey, and C. R. Woese. 1996. The Ribosomal Database Project (RDP). Nucleic Acids Res. **24**:82–85.

Meyer, F., H. J. Schmidt, E. Plumper, A. Hasilik, G. Mersmann, H. E. Meyer, A. Engstrom, and K. Heckmann. 1991. UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*. Proc. Natl. Acad. Sci. USA **88**: 3758–3761.

Ohama, T., T. Suzuki, M. Mori, S. Osawa, T. Ueda, K. Watanabe, and T. Nakase. 1993. Non-universal decoding of the leucine codon CUG in several *Candida* species. Nucleic Acids Res. **21**:4039–4045.

Osawa, S., T. H. Jukes, K. Watanabe, and A. Muto. 1992. Recent evidence for evolution of the genetic code. Microbiol. Rev. **56**:229–264.

Poynton, S. L., W. Fraser, R. Francis-Floyd, P. Rutledge, P. Reed, and T. A. Nerad. 1995. *Spironucleus vortens* n. sp. from the freshwater angelfish *Pterophyllum scalare*: morphology and culture. J. Eukaryot. Microbiol. **42**:731–742.

Rozario, C., L. Morin, A. J. Roger, M. W. Smith, and M. Müller. 1996. Primary structure and phylogenetic relationships of glyceraldehyde-3-phosphate dehydrogenase genes of free-living and parasitic diplomonad flagelates. J. Eukaryot. Microbiol. **43**:330–340.

Schneider, S. U., and E. J. de Groot. 1991. Sequences of two rbcS cDNA clones of *Batophora oerstedii*: structural and evolutionary considerations. Curr. Genet. **20**:173–175.

Schneider, S. U., M. B. Leible, and X.-P. Yang. 1989. Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase-oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage. Mol. Gen. Genet. **218**:445–452.

Schultz, D. W., and M. Yarus. 1994. Transfer RNA mutation and the malleability of the genetic code. J. Mol. Biol. **235**:1377–1380.

Siddall, M. E., H. Hong, and S. S. Dresser. 1992. Phylogenetic analysis of the Diplomonadida (Wenyon, 1926) Brugerolle, 1975: evidence for heterochrony in protozoa and against *Giardia lamblia* as a "missing link." J. Protozool. **39**:361–367.

Swofford, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Version 3.1.1. Illinois Natural History Survey, Champaign.

Upcroft, J. A., A. Healéy, R. Michell, P. F. L. Boreham, and P. Upcroft. 1990. Antigen expression from the ribosomal DNA repeat of *Giardia intestinalis*. Nucleic Acids Res. **18**:7077–7081.

van Keulen, H., R. R. Gutell, M. A. Gates, S. R. Campbell, S. L. Erlandsen, E. L. Jarroll, J. Kulda, and E. A. Meyer. 1993. Unique phylogenetic position of Diplomonadida based on the complete small subunit ribosomal RNA sequence of *Giardia ardeae, G. muris, G. duodenalis* and *Hexamita* sp. FASEB J. **7**:223–231.

Vickerman, K. 1990. Phylum Zoomastigina class Diplomonadida. Pp. 200–210 *in* L. Margulis, J. O. Corliss, M. Melkonian, and D. J. Chapman, eds. Handbook of Protoctista. Jones and Bartlett, Boston, Mass.

Dan Graur, reviewing editor

Accepted May 15, 1997