Volume 12 Number 19 1984 Nucleic Acids Research

The sequence of the gene for cytochrome *c* oxidase subunit I, a frameshift containing gene for cytochrome *c* oxidase subunit II and seven unassigned reading frames in *Trypanosoma brucei* mitochondrial maxi-circle DNA

Lambert A.M.Hensgens+, Just Brakenhoff, Berend F.De Vries, Paul Sloof, Marijke C.Tromp*, Jacques H.Van Boom* and Rob Benne§

Section for Gene Structure and Expression, Laboratory of Biochemistry, University of Amsterdam, P.O. Box 60.000, 1005 GA Amsterdam, and *Organic Chemistry Laboratory, Gorlaeus Laboratories, P.O.Box 9502, 2300 RA Leiden, The Netherlands

Received 30 July 1984; Accepted 13 September 1984

ABSTRACT

A 9.2 kb segment of the maxi-circle of Trypanosoma brucei mitochondrial DNA contains the genes for cytochrome c oxidase subunits I and II (coxI and coxII) and seven Unassigned Reading Frames ("URFs").

The genes for coxI and coxII display considerable homology at the aminoacid level (38 and 25%, respectively) to the corresponding genes in fungal and mammalian mtDNA, the only striking point of divergence being an unusually high cysteine content (about 4.5%). The reading frame coding for cytochrome c oxidase subunit II is discontinuous: the C-terminal portion of about 40 aminoacids, is present in the DNA-sequence in a -1 reading frame with respect to the N-terminal moiety.

URF5, 8 and 10, show a low but distinct homology (about 20%) to mammalian mitochondrial URF-1, 4 and 5, respectively. In URF5, the first AUG is found at codon 145, whereas extensive homology to mammalian URF-1 sequences occurs upstream of this position. The possibility exists that UUG can serve as an initiator codon.

URF7 and URF9 have a highly unusual aminoacid composition and do not possess AUG or UUG initiator codons. These URFs probably do not have a protein-coding function.

The segment does not contain conventional tRNA genes.

INTRODUCTION

Mitochondrial DNA (mtDNA) in trypanosomes possesses a highly unusual structure that is unique in nature (for reviews see refs 1-4). In T. brucei it consists of a catenated network of two types of circles, 10^4 mini-circles of 1 kb and 10^2 maxi-circles of about 20 kb. The maxi-circle contains a number of genes found in the mtDNA of other organisms (5-10) and, therefore, can be regarded as the trypanosomal equivalent of these DNAs. The role of the mini-circles is unknown. We have undertaken nucleotide sequence analysis of the maxi-circle of Trypanosoma brucei mtDNA, in an attempt to further identify the mitochondrial genes and to study their organization and mode of expression.

To date, we have reported the sequence analysis of maxi-circle segments containing the genes for the mitochondrial ribosomal RNAs of 12S

© IRL Press Limited, Oxford, England.

and 9S (6), for apocytochrome b and a number of unusual URFs (7,10). A common feature of these genes is their low degree of conservation when compared to other organisms. The ribosomal RNAs show hardly any direct homology to the ribosomal RNAs of E. coli or other mitochondria (6), although certain aspects of a possible secondary structure are reminiscent of the conserved secondary structural domains of E. coli rRNA as envisaged in refs. 11 and 12. The mitochondrial protein-coding genes also appear to conform to this pattern (7,10). The gene for apocytochrome b is only 25% homologous at the aminoacid level to its mammalian counterpart, whereas the yeast and mammalian apocytochrome b genes are about 45% homologous, indicating a larger evolutionary distance between mammalian and trypanosome mitochondria than between mammalian and yeast mitochondria.

The maxi-circle also contains long open reading frames without an AUG codon in the N-terminal moiety (see ref. 7). These occur in areas which are abundantly transcribed, but the aminoacid composition of the proteins they encode is highly unusual. This makes assessment of their role somewhat problematic. As yet, no conventional tRNA genes have been found.

In this report we present the sequence of a 9.2 kb segment on which two familiar mitochondrial protein genes (the genes for cytochrome c oxidase subunit I and II, coxI and coxII) and a number of URFs are localized. Some aspects of the trypanosomal mitochondrial genes and their organization are discussed in more detail, now that about 70% of the maxi-circle has been sequenced.

MATERIALS AND METHODS

Materials

Restriction endonucleases were from New England Biolabs or Boehringer Mannheim; DNA polymerase (large fragment), calf intestine phosphatase and T4 DNA ligase from Boehringer Mannheim; Exonuclease Bal-31 from New England Biolabs or Bethesda Laboratories; low melting agarose from Bethesda Research Laboratories; S_1 nuclease from Sigma.

DNA and assays

The isolation of trypanosome mtDNA (T. brucei 427, culture and bloodstream form) was performed as described in ref. 5. DNA was stored at -20°C as an ethanol precipitate. Plasmid DNA and M13 RF DNA were isolated according to Birnboim and Doly (13).

Restriction enzyme digestion, agarose gel electrophoresis, blot analysis of DNA fragments, nick translation and hybridization was performed as in refs 5-7. Bal-31 digestion was performed at 30°C for varying periods of time; routinely 0.5 U of Bal-31 was used per μg of DNA. Incubations were stopped by the addition of phenol.

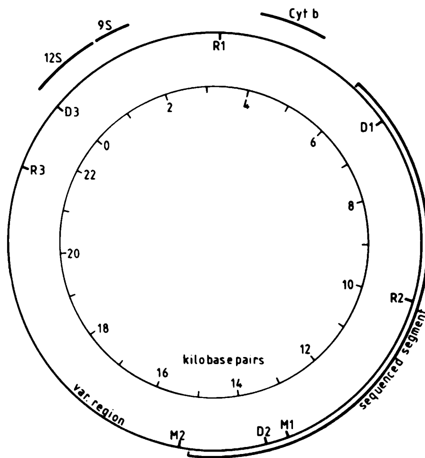

Figure 1
Partial map of T. brucei 427 maxi-circle DNA. The position of the 12S and 9S rRNA genes (6), the apocytochrome b gene (cyt.b) (7) and the variable region (var.region) is indicated together with that of a few relevant restriction sites (5), (R = EcoRI, D = HindIII, M = MboII) and the area of which the sequence is reported in this paper.

Cloning in M13 and sequence analysis

Four restriction fragments of bloodstream form T. brucei 427 maxi-circle were cloned in M13 mp8 and mp9: R_1-D_1, D_1-R_2, R_2-D_2 and M_1-M_2. Fig. 1 shows the position of these fragments relative to the position of the genes for the rRNAs, the gene for apocytochrome b and the region which varies in size in closely related T. brucei stocks (variable region, 14). The M_1-M_2 fragment was cloned in the HindII site of mp9 after blunt-ending the MboII sites with DNA polymerase I, large fragment. The nucleotide sequence of the fragments was determined using non-random cloning procedures with the use of exonuclease Bal-31 as described by Poncz et al. (15).

In a previous report we have given a detailed description of the use of this method in the sequence analysis of part of the R_1-D_1 segment (7). The procedure yields a large series of nested fragments in two orientations, with the part progessively shortened by Bal-31 oriented towards the vector's priming site. Some parts of the sequence were verified with the use of M13 recombinant DNA from clonebanks derived from maxi-circle DNA restricted with MboI, AluI or MboII, cloned in the BamHI site of M13 mp9 (MboI fragments) or the HindII site of this vector (AluI and blunted MboII fragments). Phage DNA obtained from these banks was also used to sequence across the D_1, R_2 and D_2 sites. Whenever a certain area could not be sequenced without ambiguity with clones from the Bal-31 and restriction enzyme banks, sequences were obtained with the use of synthetic oligonucleotides (prepared as in ref.16). These were utilized to prime complementary strand synthesis on M13 DNA with large maxi-circle inserts (R_1-D_1, D_1-R_2 etc.). In areas of special interest (see figs. 4 and 5) this approach was followed to thoroughly check the obtained sequences of bloodstream-form maxi-circle DNA and to compare them with the nucleotide sequence of culture-form maxi-circle DNA, which was cloned as a EcoRI x HindIII digest in M13 mp8 and mp9. The colinearity of the cloned M13 inserts with maxi-circle DNA was checked with S_1 nuclease analysis, as described in ref. 15.

DNA sequence analysis was carried out by the dideoxy nucleotide chain-termination technique according to the method of Sanger et al. (17). All DNA fragments were completely sequenced in both directions. The nature of the procedures followed, provides an ample source of overlapping clones. Each part of the sequence is derived from at least two

```
<--- URF4 ends at -401
   1 TATATAATAA AAAAATAGTA TATAATAATA AGTAATACTA AACTTATACT ATAAATTAAG TGAAAATTTA AATATAAATA AAAGATATAA TTTTTTGTTG
 101 AAATAAATAT TAGGAATAAA AAGCAAAAAT TATTCACACT TAACACAAAT AGTAAACTAA CGATAGCAAA GCTGTTTAAT CCAATTAAAA CGCATGTACA
 201 AGATTGAAAT AATAGAAGTT TGATGAATAA AATATAAAAA TAAATGAAGC TAATTAGTAG AATTATTAAT ATAAAACAAA ACAAAATATA AAAAGTTAAC
 301 ATATAAATAA AAATAAAGAC ACCAAGTCTA ATATAAAGTT GCTCCATAAA CAAAATTAAA AAGGCGATGT ATAATTTGAA TAAAATTAAT AATGTGTAAA
 401 ATAGGCATAA AATTCCAAGT CATTCTTCAT CAAAAACTAA AAAACAAAAA TCACATAGGA AAAAACAGTA GTTTAATATC ATAAAATATA ATAATATAAA
 501 TAATAATATA AAATTTATTA AGTTTAACAT GTAGTAATAT CATAGAACTA AAATTTTATA TCCAAATCTA CTGGACATTA ATAATAAAAA GAGCAATAAG
 601 CTAAATATTT CAAAGAGGAT TGATATAATA ATAATATGAT TAATAAATAT AAATAAGAAT ATAATAATGT ATTGAATAAT AATAATAATG AATAAAAATC
 701 TGGTATCGAA TGATAGAAAG CAAAAAAATA ATGTAAAGCA AAATAAGAAT AAGAGTATAA AGATGAAACA AATATAAGAA TCTAATAATG TTATTCAAAA
 801 TAGGTTAATA ATTAATAATC AGAGTAAATC AAAGCTTAGT AATGTTAGTG TAGTATAATC ACATAAGATA ATAAAGCTGT AGATAATAAG AAATATAAAT
                                   <---| Start URF4
 901 ATGTGTATGA TATATAAAAA CAAGGATTTT TTGGGGGTTT AGGGACAGAG GGTTTATTTT TGAGGATTTT AGGAGGAGAA AAGGGATGGG AAACAGAAGG
                                |<--- End URF5
1001 ACATAAGAAA AGTTTCGTTA TTAGATTAAA AAAGTATGCA AATAATTTTT GTAATAGCAA TAAATGAAAA ATTAATGAAT CCCATTGTAA ATAAAAAAAG
1101 TAATATAAAT GTTTGTGCAG TTGTAATTTT TAATCTACAG CATATAACAC GTGGTATAAG AAAACCTAGA ATTAGTATAA GAATAGATTT AAAGCATATA
1201 AAAAGGCCAC CAAAACATAA GCAGCTAAAT AATATAGTAG TTAGTAATAA ATGATTAATT TCAAGGACGG AGTATATGAC AAAAAAAATA CCAGATAATT
1301 CAGTAACAAG GCCAGCAACA AGTTCACTTT CACATTCTAG ATAATCAAAG GGTAAACGTA ATCCATCAAG AAGCAGTCCA ATTCAAAATA GACAAATGAA
1401 TAAAAGACCA AGTATAAAGC AATTTTGTAA AGAAAGTTGA CTTATACAAA TATCTTTTAT GCCAAAGAAA CAAAAGTAGT CTAGTATGTA TATACAATAA
1501 AGTAGAATTA AAATGGAACA TTCAGATAGA ATACTAAAAA ATAAAGTTCT CATGGCAGCT AGGTAAATAA AGCAGCTTGA AAATAAAAAA CAACCAACGA
1601 AAAAAATGCA AAATACATTA GAAAATAAAT GAAATCCTAA TAAAAAAAGA AGTGTAAAAC CTTTGTCAAA TATTATTATG AATCCAAGTG GGAAAAAGAA
1701 CCAAGGAAAA AAAATACAAA AAGCTGTAAT GAATAAACTA GATATAAATA ATATAGAGTC AACACCAATC ACAAATAATG TAAATTTAAC AATAATTTA
1801 ACTCCATCAG TAATAGGAGT AAGAAGACCA AATAAAAAAA GTGCAGGTCC TATTCTGAAT TGCACAATAG CTAAAATTTT ACGTTCACAT AAACTAACAT
                                                          Start CoxII |---> <---|Start URF5
1901 ACCCACATAA GACAGATAAA ACGAGTATAA ATATAACAAT AAGTATGCAT ATATCTAAAT GTAATAATAA CAAATGAGTT TTATATTAAC TTTTTGAATG
2001 ATATTTTTAA TGGATTCAAT AATTGTATTA ATATCTTTTT CAATATTTCT ATCTGTATGA ATATGTGCAT TGATTATAGC AACAGTATTA ACTGTAACAA
2101 AAATAAATAA TATATATTGT ACATGAGATT TTATATCATC AAAATTTATA GATACATATT GGTTTGTACT TGGAATGATG TTTATATTGT GTTTATTGTT
2201 AAGGTTGTGT TTGTTGTTGT ATTTTAGTTG TATAAATTTT GTGAGTTTTG ATTTGTGTAA AGTAATAGGT TTTCAGTGAT ATTGGGTATA TTTTTTATTT
2301 GGAGAAACCA CGATATTTAG TAATTTAATA TTAGAAAGTG ATTATTTAAT AGGAGATTTA AGAATATTAC AGTGTAACCA TGTATTGACA TTGTTAAGTT
2401 TGGTTATTTA TAAATTATGA GTATCTGCAG TAGATGTAAT ACACTCATTT ACAATATCAA GTTTAGGTAT AAAAGTAGAG AACCTGGTAG GTGTAATGAA
                                                                                              End COXII --->|
2501 ATAATTTTGT TTGCTACAAA TAACGCAACT CTTTACGGAC AATGTAGTGA ATTGTGTGGT GTATTACACG GTTTTATGCC TATTGTAATA AATTTTATAT
                                Start URF6 |--->
2601 AGAAAGGTAT ATAATCTATA ATGAAAGGGG ATTTTAAGAT TGGCTTTGAT TGAGTCGTGT TTTTGATTTG TTATGTATTA GAACATATGA TTTTATATTG
2701 TGATGGTTTG ATCTAGATTT TATATTATAT GATTTTGTAT TCGATTTTGT TGTATGTATT ACATTTATAT TTATATTTGT TTTAGGTTTT TTTATTAGAA
2801 TTTTTTTTAG TTTTGTGTTT GTATTGTTAT TTATAACATT TTTTGGAATT TGTTCATTAA CAATGTTATT TACAGGGTAT TATATATATT ATATATATAT
2901 ATTATATAAT TTTATATGTT TTTTTTTTGC ATTTGGTATA AATTTTTTGA TATATTATAT CGAGTTTTTC ATATTTATAA CATTCCATAT ATTTTTCGAT
3001 TTTATAAGTT TTTCTAATTA TATATATAAT TATTTTGGAA TATTGTATAT GTTTAATGTA ATGTTTTGTG CATATTTATT TTGTTTATTT TATTTTGTGA
3101 TATATTTTTT ATTTTGTTTT ATATTTTTTG TAATACGATG TTTATTTATA GTAATAATGG ATTTTTTATT TTTTAATTTT GATATATTTG TATCTATATT
3201 ATTATGTGAT ATAGTATATT TAGATTTTAT AAGTTTATTA TTATTATATT TTAATTTTAT ATTTAATTTT ATTTATGGAT TTTTTAGTTT TGTGATAATT
3301 TTAGGTTTAT TATTTTTATT GTTATTTTTA GTAATAAATT TATTTTTTGG ATTTACATTT TTAGTATATG GTATACAAAT CATATTATTA TATTATGTAT
3401 ATTGATTATA TATGATATAT AGTAGAAGTT GTTATATATT GATGCCAGCA ATATTAATAT TTTTTAAGTT TATATATTTT GATGTATTCT TTGTGTTTGT
3501 ATTTATTTTA ATTTTATTTA TTATATCATT TTTTAGTTTT TTTTTAAAAG ATTTTTTATT TTTATCATTA TATTTTGATA TATTCGGATC ATTATATAAT
                                                                                |<--- End COXI
                                                                     End URF6 --->|
3601 TACGATATAT TATCATATAG TATATTTTAT TATCAAAATA ATCAGTTTTG TTTAACACAG TTATTATCAA TTTATATATA AAAGAATAAT AGGAAGGCTT
3701 GTCAAAAAAA AATATAAAAT ACGAAGCAAA AGCAAAGTAT TATTAATAAA TAGTCAAGTA TTATATGCGC AAAATCAATA ACAAGTAAAT AAATAGCCAT
3801 ACAAACTGGT ACTCATGTAT AAAAATAGAA GAATATAGAT AAGCTATATG TAAACAAATT AATAAAAAAT AAACAGTAAT CTCAGAATAA TATAACATTA
3901 AACAAACAAC AGCAAAATAT AACTAAAAAA GTCAACAAAA GCATACCATA CAATGTAAAC GCACTTCAAA ATAAAAAGCT GATAGGATAA TCAGAAATTC
4001 TTCTTGGAAA AGCAAACATT CCTAAGCTAT GTAAAGGAAA AAAAACCATA TTAGAACCAA ATCATAATGT GGAAATGAAG AAAAATAATC AAAATGTGTG
4101 AAGTTCAATA GGAATTCATT TCATGAGAAA ATGAAAAAAC CCACCAAAAA CCCCTACAAC AGCACCAAGT GATAAAACAT AGTGGAAATG TGCAACAACA
4201 AAATAAGTAT CATGCATCAA AATATCAATA CCAACATTTG ATAGAAATAA GCCAGTTAAT CCACCAGCAA GAAACATAAG TATAAACATA TATATAAAAT
4301 AAATTTCAAA ACAAATACAC ATATCTGTGA ATAAGAAGCT ATAGATTCAA TTAAATAATT TAATGCATGT AGGTAAGCCT ATTAGTACAG TAATACTTCC
4401 AAAATAAGCT CTAGAATCAA CATCCATACC AACAACAAAC ATGTGATGCG CTCAAACAAA CATACCTAAT ACAGATATCA GTAGCATAGA ATAAATCATA
4501 GCGACGGAAC TAAAAACGCA TCTAAAACTT GTAACTTCAA TAATAGTGGA AACTAATCCA AATACAGGTA GAATTATGAT ATAAACCTCT GGATGTCCAA
4601 AGAACCAAAA TAAGTGTTGA AATAAAACTA GATCTCCACC TCCAACAACA TCATAAAATG ATGTGTTAAA ATTTCTATCG CATAATAATA ATGTAACTCC
4701 ACCAGCCAAA ACGGGAAGTG TAATAATCAA AAGTATGGAT GTCAACAAGG CACCTCAAAT AAAAAGTGTT CAAATAAGAA AGCTAAAATA TTTTCGTCTA
4801 CAGCAAAAAA TTGTACCTAC AACATTAATG GAATTTAATA TACTAGATAT ACCAAGAAAA TGCACAGAGA AAATGATAAA GTCACATGCT AAGCTAGAAT
4901 GAAAATCAAT ACAAATCAAC GTAGGATACA AGGTCCAACC AACCCCCATA CCCTCTTCAG TCAAAAACCC ACTTACTAAG CAACCAAATC CTCCAATAAA
5001 CATTCAAAAG CTCATATTAT TTATACGAGG AAATACCATA TCGGGAAACC CAACCATGAC GGGCGCAAAA TAGTTTGTAA ACCCACCCAT AGTTATAGGC
5101 ATAATAAAAG CAAAAACCAT AATCAACCCA TGTGAAGTAA TGAGTACGTT GTAAAACTGA TAATCACCAA ATAAAACTCC ACAACCAATT AGAGAAAGTT
                                                   Start URF7 |--->
5201 CTAATCTAAT AAATAAGGAG TAAATATAAC CGATAAATCC ACATAAAATA GCAACTAAAA GATAACAAAT TCCAATCATT TTATGAGAAA CACTTAAGCA
                    <---| Start COXI
5301 CACAAGACAT AGAAAAAACA TTTTATAAAA CCAAAATAAA AACAAAACTA TTTCTTCGCA ATTATAAACC AAAATTCCAT CAACCCCCTC TCCCCCTCCT
5401 CCCCACAAAA CCCCACACTT TCTCCCCAAC TAAAACCCCC CAATCAAACT CCCCTTCAAT TTCGTCCTCT TCTAAACCCA ACTTCCCCAA ATCCCCCTTC
5501 CTCTCCCCCA CAAACCCCCC TTTCCTCATA GATCAATCCC AACCAAACTT AAAGCCCCTT TTTAATAAAT ACACAAAGAT AACAATAAAT TATAAGAAGG
                         Start URF8 |--->
                        End URF7 --->|
5601 AAATTTATAG AAAGCACAAA AATAAAATTA AATTAGAGTA ATTGAATGTT AAAATTAAAT TTAATATGTA TAAATTTTAT ATTGTTAATT GTTACAATAA
5701 TATATATATA TATAAACTAT AGTTTTTGTA TTGGAATAGA AATCAATTAT GTATATGTAA ATATATATTT AAATTACATC AGTCTATGAT TTGTATTTTT
5801 TATGGGAATT ATTATGTACA TATTAATATT TTTATTATCA AAGAAATGTG TATCATATAA TAAATATTTT TACATAGTAA TGATATATAT GTATATATAT
```

```
5901  ATTAATGTAG TGTTAATAAT AATATTAGAT GATTTTATGT GTTTTATGAT AGCCTTCGAA AGTCTATTTT TCCCTATATG TCTAGTAAGT TTATTTTTTA
6001  ATTTTAATAA TAGATTTATT TTTGCTATAT TCTATCTTAT AATATTTAGT TCAGTTAGTT CAGTGGTATG TATAATTATA TGTATAATAG TAATATCTCA
6101  TTTCAACATT ATAAATTTAC AGGCTTTTAT TGATGTATGT TATTTTGATA GTTTGTATTC GGCAATTTTT ATATGAATAT TATTATTTAT AATGTTCGCT
6201  ATAAAATACC CAATCTGACC ATTCCATGTG TGACTACCAG AGATGCATGT AGAGGTAAAT ACAGAAATGA GTGTTTTATT AGCAAGTATT GTGCTGAAAA
6301  TAGGTTTTTT TGGTGTATAC AAATTTTTAT TTATCGCATT TAATACGATA TCAATATGAT TTTTAGGTTT TATAGATAGT GTAATTGTGT TGGGTTTAGT
6401  ATTTATAGCA ATGTCACTAA TATTTTTATC AGACTACAAG AAAATAATAG CGAATTGATC AATAATACAC ACGGGTATAG GATTAATATT ATTATGACAT
6501  AATGACATTT TGTTTGTAGG TTTACTAATA TTATGTAATC TAGCACATAT ACTAAGTTCA TCCTTTATGT TTATTGTAAT AGGATATATG TACGACAATT
6601  ATGGTGTAAG AATTTTTTTA TTGTTAATTT CATTTTTTGG TATTAGTATA TGAAGTTCAT TGTTTTTATG TTTATTTTTA TTTAATATAG ATTTCCCGTT
6701  TATGTTATTA TTTTATGTAG ATATATTTAT TTTGTATGGT TTGATATCTA TATCATTTAT ATATATAATA AGTTTTTATA TAATAACTTT AACGATATTT
6801  TTATCATCAA TATACATCTA TATGTGTTTA AGTTTTTATT CATTTGTATG GTTGGATAAA TATCTTAGAC TTGATGTTAG TATAAATGAT ATATATGTAT
                                               |<--- End URF9
                                                    End URF8 --->|
6901  TTATGTCAAT ATCAATATCA ACTATAGTAT TTTATTATTT TATATATTTA TTAATATAAT ATGTATAATA CAACAAACAA ATCTCTTTAC CCCCTTCAGT
7001  GATCCCTCCC CATCAAAACT TCTCCCCCCA AAACCCATCT CCCATTCACC CCAAACCTAT GGTTTCTCCA ACACTCCATT CCTGTTCACA CCGTGATTCT
7101  TCTCAACCCC GCCCCCCGCT CTGCTCTCTC CTTTTAAAAT CCCTAATACA CTTTTGATAA CAAACTAAAG TAAAAAGGCG AGGATTTTTT GAGTGGGACT
7201  GGAGAGAAAG AGCCGTTCGA GCCCAGCCGG AACCGACGGA GAGCTTCTTT TGAATAAAAG GGAGGCGGGG AGGAGAGTTT CAAAAAGATT TGGGTGGGGG
                                                   Start URF10 |--->
7301  GAACCCTTTG TTTTGGTTAA AGAAACATCG TTTAGAAGAG ATTTTAGAAT AAGATATGTT TTTAATATTT TTTTTATTTT TTATAATGTT TGGGTTTATA
7401  TCAGGTTCAT TTATGTTTGG TAGGAATTTT CTAAGTTTTT GATTATCTTT AGTAATGATA ATATTTATTG TATTGTGTAT GATATTTAGT TTTTTAATGG
                     <---| Start URF9
7501  TATCAGTATG TTTATACGGT TATTATTATT ACGATTTTTG TTTAATACTA ATGTTAGATT TTTGTTTTAT ATGATTAACA TACGTATGTT CAGGTTTTTA
7601  TATGTTTATA ATGTTATTGA TAAATATGGT ATTTTGTTTT ATAGTATTTT ACGCATTTTA TTATATGTAT TTTGATATGT TGTTAGGGCG TTTTTTGATT
7701  ATATTTTGAA TATTTGTTGT GTGCATGAAT TTATTCATCC TATCATATGA TTTTTTAACA GCTTACTGCG GATGAGAATT ATTAGGGTTA TTCTCATTTT
7801  TTTTAATTTC ATATTTCTGA TACCGTTTTT TTGCATTAAA ATTTGGTTTT AAAGCTTTTT TTATAGGTAA AATAGGAGAT GTGTTATTAA TATTCGCTTT
7901  TTCTATAATA TTTTTATCAA ATGGTTTTTG TATGACAACT TTTTATTTTT TAAATTTTTT TTGTATGGAT TATTATTATA TAGAATTTTC TATATGTTTG
8001  TTAGTAGGAT GTGCGTTCAC AAAAAGTACA CAATTCGGCT TACATATATG ATTACCAGAT GCTATGGAAG GACCTATCCC AGTATCAGCA TTAATACACG
8101  CAGCTACATT AGTTGTTTGT GGAATAATAT TATTAAGTTT TGTTTATTGA TGTTTTGATT TTTGATTTAG TTATTTTTAT AATTTGATAG GATGGTCTAC
8201  ATTAATTTTA ATATTAATGA CATTGTGTGT GTTTTATAAT TTTGACGTAA AACGATACGT AGCGTTCAGT ACAATATGTC AAATTAGTTT TTCTATGTTT
8301  TGTTGTCTGT GTATAGATAT ATATATAGGT AGTTTATTTT TTTGTTACCA TATGTTCTAC AAAGCAACAT TATTTATAGT ATTAGGTATA TGAATACATA
8401  TATTTTTTGG GTTACAGGAT TTAAGATGTT ATTTTTTTAT GTATTTTTGT GGTTGTGTGT TAGCGCGTTT GTTATTAATA TTCGCAATAT TAAACTCATG
8501  TTCAATTTGA TTTTTATGTG GTTTTTATTG TAAGGATATG TTATTAGCTT TATTGATGTT ATTATCATTT TATAATATAA TAGAATTTTT GTTTATAAGT
8601  ATAATTTTTA TATTTTTTAC AATGATTTAT AATTATTTTT TGTTATTTTT TTTGATGTTT GTGTTCAAAT GTTTTTGTTT GGTTGATTGT TTATTTTTAT
8701  TATTTGATTA TGAATGTTGT TTAGTATATT GTTTGATAAG TTTGTATATG TGTATTTTAA GTATATTTTT TATAATAGAT TTTGTATGTA TATTTGTATT
8801  TTCAAGTTAT TGTGTATTTT GATCATTTTT TTTAAATTTT TATAATTTTT TTGATATAGC AATTTTTGTG GTTTTTTTAA TATTATCAGT AGGATTTTTA
8901  TATTATGGTT GTTTATTTTT TTATTTTTTC AATATAGATT GCATAATGTT GTTTTGGAGA ATTTTTTTTG TAATAATAAT TTTAGTAGTA TTTATGATAT
9001  TTTGTTGTTG ATATTTTGTT TGTATGATCA TATTTATGTT ATTATTTGTA TGAAATTTTG TTATATATTT TAGATATAAT TTGAAATATT GTTTATTTTT
                  End URF10 --->|
9101  TTGTATTTTG TGAATATTGT ATGTATAAAT AGTATAATCA AAAGTAAAAA AAGTAAAGAA ACCAGATTAG ATTTGTAAAA AAGTCAAAAT ATTTTATAAT
```

Figure 2
Nucleotide sequence of a 9.2 kb maxi-circle segment of *T. brucei*. First and last nucleotide of a number of protein genes and URFs has been indicated (see also Table 1). The sequence strategy has been described under methods. Nucleotide number 1 corresponds to nucleotide 2501 in ref. 7. Genes were identified by comparison with aminoacid sequences of human (18) and yeast (19-26) mitochondrial proteins.

independent clones in each direction. In a previous paper the sequence of the first 2520 nucleotides of the R_1-D_1 fragment was reported (7). The present paper provides the sequence of the remainder of the R_1-D_1 fragment (nucleotide 2521-3332) together with that of the other fragments to a total of 9200 nucleotides (see Fig. 1).

RESULTS AND DISCUSSION

Fig. 2 presents the complete nucleotide sequence of a 9200 bp segment of the maxi-circle of *Trypanosoma brucei* (see also Fig. 1). Begin and end-point of a number of genes and unassigned reading frames are indicated. These were obtained by translating the nucleotide sequence into aminoacids with a genetic code in which only the assignment for UGA (encoding tryptophan in most mitochondrial genetic systems, including

Table 1 Mitochondrial genes and URFs on a 9.2 kb maxi-circle segment

	nucleotide coordinates of reading frame 5' — 3'	1st AUG at codon position
URF-4*	938--401	121
coxII	1971-2599	2
URF-5	1985-1029	145
URF-6	2638-3678	142
coxI	5322-3675	36
URF-7	5252-5638	-
URF-8	5637-6956	4
URF-9	7519-6938	-
URF-10	7353-9125	2

* 401 nucleotides at the 3'-end of this URF have been published in ref. 7

trypanosomes; 7) differs from the universal code. The genes and URFs found, together with their coordinates in the sequence of Fig. 2 and the position of the first AUG-codon, are listed in Table 1. The gene for cytochrome c oxidase subunit II (coxII) and the URFs 6, 7, 8 and 10 run clockwise, the gene for cytochrome c oxidase subunit I (coxI) and the other URFs, counterclockwise (see also Fig. 6). An extensive discussion of some of the characteristics of the genes and URFs is given below.

The genes for cytochrome c oxidase subunits I and II

The aminoacid sequence of the *T. brucei* coxI and coxII genes is given in Fig. 3 in a comparison to the analogous genes in yeast and human mtDNA. The coxI gene shows an overall homology with the yeast and human genes of 38%, the coxII gene of 25% assuming a limited number of insertions/deletions. This is consistent with the pattern also observed for the rRNA genes (6) and the apocytochrome b gene (7), which also show a rather low degree of direct conservation. There can be no doubt, however, that we are dealing with the coxI and coxII genes, since many of the aminoacid substitutions are conservative, which results in very similar hydrophobicity profiles for the *T. brucei* and yeast/human gene versions (plots not shown). Furthermore, some of the putatively functional aminoacids are conserved: e.g. in mammalian coxII, His_{109}, Cys_{208}, His_{216} and Met_{219} (the coordinates used are those from Fig. 3)

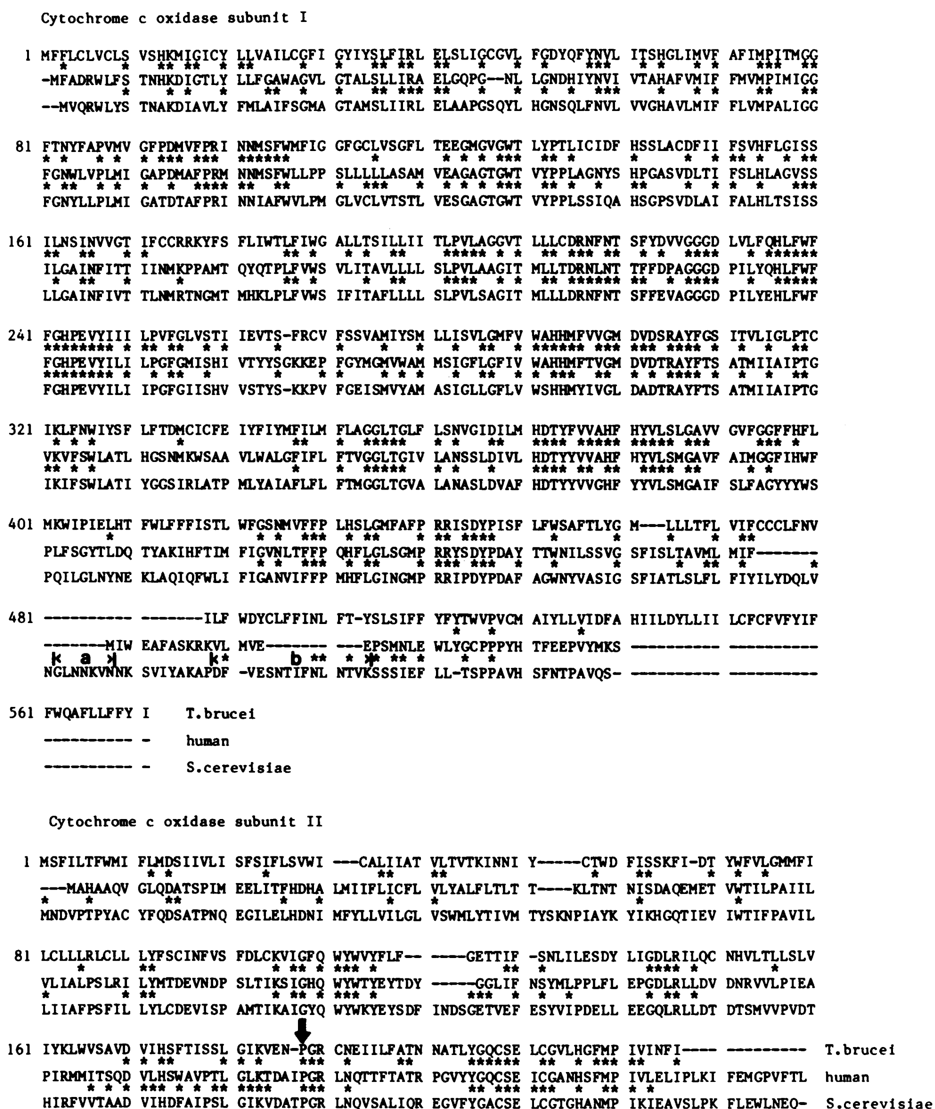

Figure 3

Sequence comparison between coxI and II genes. The cytochrome c oxidase I and II genes were lined up with the analogous genes from human (18) and S. cerevisiae (yeast) mtDNA (19,20). * indicates homology with the T. brucei sequence, - indicates the position at which a deletion is assumed. The gene sequence is presented starting with the first methionine (see Table 1). a and b in the coxI sequence indicate proposed intronic sequences in yeast coxI (19).

may serve as ligands for Cu binding (27): only His_{109} is not conserved in T. brucei. Also a stretch of aromatic aminoacids around position 112 with a possible function as transmembrane electron channel (28) is present in T. brucei. Moreover, the proposed sites for O_2-binding (position 234-249) and heme a_3 attachment (371-382) of the coxI subunit (29) are almost completely conserved in T. brucei. However, only 42 out of 94 invariant residues in human and yeast coxII and 159 out of 250 invariant residues in coxI of 5 species (29) are present in T. brucei. We anticipate, therefore, that it will be instructive to closely inspect the T. brucei mitochondrial protein sequences in order to acquire more information on the composition of the functional domains of the proteins of the respiratory chain.

Two stretches of 21 and 45 nucleotides, respectively, at the 3'-end of the yeast coxI gene were assumed to be introns to minimize the size difference between the yeast and human version of the protein (19). We find, however, 4 identical aminoacids in the T. brucei sequence and the larger yeast "intron". We have, therefore, not omitted these residues from the sequence alignment (see Fig. 3). In view of the low degree of conservation in the aminoacid sequence at the C-terminus of coxI (29), it cannot be excluded that rather large variations in size in this part of the protein are allowed.

The genetic code; the coxI and coxII genes display a high cysteine content

A striking feature of the sequences as presented in Fig. 3 is the relatively high content of cysteine residues in the T. brucei coxI and coxII genes (e.g. 22 cysteine residues in T. brucei and only 1 in human coxI). Moreover, the apocytochrome b gene and some of the URFs (see below, Fig. 5) show the same phenomenon. Since cysteine is both structurally and functionally an important aminoacid, it is somewhat surprising to see such large differences in cysteine content between analogous proteins in different organisms. The question arises, therefore, whether the UGU and/or UGC triplets, which specify cysteine in the standard code, have a different assignment in the T. brucei mitochondrial genetic system. We have checked for this reason whether the alignments as shown in Fig. 3 and in ref. 7 allow an unusual assignment for UGU and UGC. We have also included in this study some other codons of which the assignment deviates in various mitochondrial genetic systems: UGA, AGA, AGG, AUA, AUU and the CUN family (see refs. 7 and 30 for a more extensive discussion).

On the basis of previous sequence analysis of trypanosomal mtDNA

Table 2 The genetic code in T. brucei mitochondria

Codon	Number	Aligned with aminoacid	H A(%)	H B(%)	Y A(%)	Y B(%)
AUA	71	Ile	30	8.5	24	9.9
		Leu + Val	30	18.7	25	19.6
		Met	5.6	4.9	4.2	3.9
AUU	49	Ile	25	8.5	27	9.9
		Leu + Val	29	18.7	33	19.6
		Met	2.1	4.9	2.1	3.9
AGA	13	Arg	62	1.8	69	2.0
		Lys	23	1.8	15.4	2.0
		Ser	7.7	6.1	0	7.2
AGG	4	Arg	25	1.8	25	2.0
UGU	38	random with 17 different aminoacids				
UGC	9	random with 9 different aminoacids				

The data were compiled from the gene for apocytochrome b (7), coxI and coxII. Comparison was made with human (H) and yeast (Y) mitochondrial gene sequences (18-20,24). The frequency A at which T. brucei mitochondrial gene codons line up with a certain aminoacid in the human and yeast mitochondrial protein sequences is given compared to the frequency B at which that particular aminoacid occurs in those sequences.

(7), we were able to assign UGA and the CUN codon family to tryptophan and leucine, respectively. This assignment is confirmed by data derived from the coxI and II genes.

The data for the other codons are presented in Table 2: No consistent pattern could be observed in the alignment of human and yeast coxI, coxII and apocytochrome b aminoacids with T. brucei UGC and UGU codons. The residues found more or less reflect the composition of the proteins studied. Furthermore, two cysteine residues conserved in human and yeast coxII are also encoded by UGU. The high cysteine content may therefore be a real feature of trypanosomal mitochondrially encoded proteins. This raises the intriguing question how oxidation of these proteins is prevented, particularly in the case of coxI, which contains the oxygen binding site of the cytochrome c oxidase complex (see ref. 29).

Also the assignment for the AUA/AUU and AGA/AGG codons does not deviate from the universal code in trypanosome mitochondria. AUA/AUU line up predominantly with isoleucine (and closely related aminoacids such as leucine and valine), and not with methionine, which is specified

by these codons in mammalian (18) and insect mitochondria (31), AGA clearly codes for arginine and not for a stop (as in mammals) or for serine (as in insects). The assignment for AGG is less firm, due to the low number of AGG codons in the genes studied sofar, but also in this case the universal code appears to be followed.

The gene for cytochrome c oxidase subunit II contains a -1 frameshift

The nucleotide sequence as presented in Fig. 2 indicates that the reading frame for the coxII gene is not continuous: a -1 frameshift has to be introduced to link the C-terminal 39 aminoacid residues to the N-terminal moiety of the protein. This shift should occur in a rather small area (around residue 188, the arrow in Fig. 3) as judged from the position of large homology blocks that flank this residue on either side in the two different frames. Repeated sequence analysis in two directions, also including the use of ITP to reduce compression of bands, with clones from different banks prepared with different batches of maxi-circle DNA confirms the sequence, which virtually rules out possible chance of sequence errors and trivial cloning artifacts (results not shown).

The DNA used in this study, however, is mtDNA from the bloodstream form of T. brucei 427, a strain that has been cultivated with the use of laboratory animals ever since its isolation from sheep, approximately 10 years ago. A functional respiratory chain and Krebs cycle are absent from bloodstream T. brucei (32) and it is conceivable that a silencing mutation in the coxII gene could have occurred. We have performed, therefore, sequence analysis of the coxII region of the maxi-circle of cultured T. brucei 427, in which the respiratory chain is fully operative (32). The result of such an analysis is shown in Fig. 4, which gives the sequence of the relevant area flanked by a stop codon in either frame. Both the sequences shown (A, standard procedure, B, procedure with ITP) perfectly match that of Fig. 2. Moreover, a similar analysis of the sequence of the opposite DNA strand and of other areas from the D_1-R_2 fragment (to a total of about 1650 nucleotides) did not reveal any difference with the sequence as given in Fig. 2. The gene for cytochrome c oxidase subunit II, therefore, is also discontinuous in respiring trypanosomes. In order to explain this phenomenon, a number of possibilities might be considered:

1) The coxII gene is a pseudo gene, whose function has been taken over by a copy residing in the nucleus. We consider this unlikely since

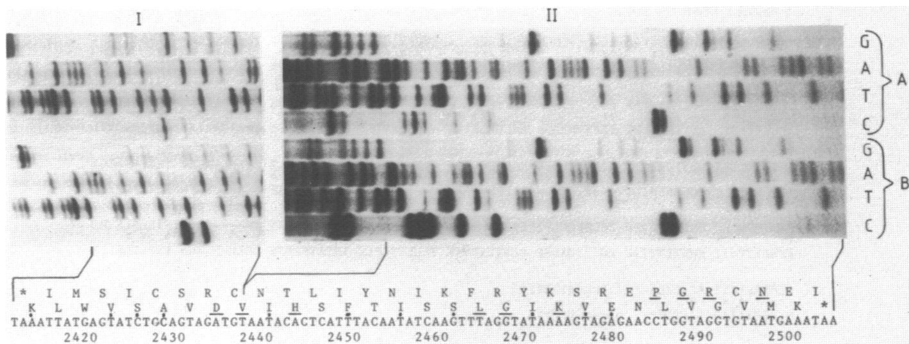

Figure 4
Nucleotide sequence of part of the cytochrome c oxidase subunit II gene from cultured T. brucei 427. M13 mp9 DNA containing the D_1-R_2 insert was submitted to sequence analysis utilizing a synthetic oligonucleotide primer. The figure shows relevant parts of an autoradiogram of a 2 hr (II) and a 3½ hr (I) run of a sequence reaction mixture obtained from D_1-R_2 in mp9 primed with 5'-CCA.CAC.AAT.TCA.CTA.CAT.TG-3'. This oligonucleotide is complementary to nucleotides 2540-2559 of the sequence of Fig. 2 (which correspond to residues 197-202 of the aminoacid sequence of Fig. 3). The reactions were performed under standard conditions (A) or with replacement of GTP by ITP (B). The obtained nucleotide sequence is outlined underneath the autoradiogram together with the aminoacid sequence in two different reading frames. Amino acids which occur at the analogous position of both yeast and human coxII are underlined, whereas conservative aminoacid substitutions are dotted.

no cross-hybridizing bands appear on blots of restriction digests of nuclear DNA from dyskinetoplastic trypanosomes which do not have maxi-circles (such as T. evansi, see ref. 33). Similar experiments with T. brucei nuclear DNA reveal no bands other than those attributable to contaminating maxi-circle DNA (results not shown).

2) Cytochrome c oxidase is dispensable also in cultured T. brucei. A branched electron transport system with cytochromes a_3 and o as independent alternative oxidases has been proposed for kinetoplastida such as T. mega (34) and C. fasciculata (35). Evidence of cyanide insensitive terminal oxidases has also been obtained in cultured T. brucei (36). The possibility exists, therefore, that in cultured trypanosomes electron transfer to oxygen can proceed, at least to some extent, without the involvement of cytochrome c oxidase. The inactivation of this oxidase in the T. brucei strain studied, however, must have been a very recent event, as judged from the fact that the mitochondrial encoded cox genes are still highly conserved and only one apparent gene silencing mutation has occurred. Most likely, such a cox^- strain would no longer be viable

A

```
    T.brucei URF5

  1 -YKTHLLLLH LDICILIVIF ILVLSVLCGY VSLCERKILA IVQFRIGPAL -FLFGLLTPI TDGVKLFVKF TLFVIGVDSI
         ****      ***     *          * *****    * * **      *** *   *  *** *   *       *
    MPMANLLLLI --VPILIAMA FL-------- -MLTERKILG YMQLRKGPNV VGPYGLLQPF ADAMKLFTKE PL--KPATST

 81 LFISSLFITA FCIFFPWFFF PLGFIIIFDK -GFTLLFLLG FHLFSNVFCI FFVGCFLFSS CFIYLAAMRT LFFSILSECS
            *             **              *** *        *  *    *    *          * -
    ITLYITAPTL ALTIALLLWT PLPMPNPLVN LNLGLLFILA -TSSLAVYSI LWSGWASNSN YALIGALRAV AQTISYEVTL

161 ILILLYCIYI LDYFCFFGIK DICISQLSLQ NCFILGLLFI CLFWIGLLLD GLRLPFDYLE CESELVAGLV TEL----SGI
     ***          *          *  *         *    *    *     * ***  *  ***** *    *
    AIILLSTLLM SGSFNLSTLI ---TTQEHLW LLLPSWPLAM MWFI-STLAE TNRTPFDLAE GESELVSGFN IEYAAGPFAL

241 FFVIYSVLEI NHLLLTTILF SCLCFGGLFI CFKSI----- -LILILGFLI PRVICCRLKI T-----TAQT FILLFLFTMG
    **       *     *** *        *                * *   **   *    *            *  *  *
    FFMAEYTNII MMNTLTTTIF LGTTYDALSP ELYTTYFVTK TLLLTSLFLW IRTAYPRFRY DQLMHLLWKN FLPLTLALLM

321 FINFSFIAIT KIICILF    T.brucei URF-5
          * *   *
    WYVSMPITIS SIPPQT-    human URF-1
```

```
    T.brucei URF8

  1 --SNWMLKL- ---------- -------NLI CINFILLIVT IIYIYI-NYS FCIGIEINYV YVNIYLNYIS LWFVFFMGII
        ****                      *       **   **     *                      *    *
    PKPTRMLKLI VPTIMLLPLT WLSKKHMIWI NTTTHSLIIS IIPLLFFNQI NNNLFSCSPT FSSDPLTTPL LMLTTWLLPL

 81 MYILIFLLSK KCVSYNKYFY IVMIYMYIYI NVVLIIILDD FMCFMIAFES LFFPICLVSL FFMFNNRFIF AIFYLIIFSS
                     *  *   *                   * * * **     *                 *  *
    TIMASQRHLS SEPLSRKKLY LSMLISLQIS LIMTFTATEL IM-FYIFFET TLIPTLAIIT RWGMQPERLN AGTYFLFYTL

161 VSSVVCIIIC IIVISHFNII NLQAFIDVCY FDSLYSAIFI WILLFIM-FA IKYPIWPFHV WLPEMHVEVN TEMSVLLASI
    * *    *   *          *            *   *      *   * *  * *    *  ***  ***      *  **
    VGSLPLLIAL IYTHNTLGSL NILLLTLTAQ ELSNSWANNL MWLAYTMAFM VKMPLYGLHL WLPKAHVEAP IAGSMVLAAV

241 VLKIGFFGVY KFLFIAFNTI SIWFLGFIDS VIVLGLVFIA MSLIFL--SD YKKIIANWSI IH-TGIGLIL LWHNDILFVG
     ** *  *      *            *    *  *    *   * * *  *  *   **  **  *        *      * *
    LLKLGGYGMM RLTLILNPLT KHMAYPFL-- VLSLW-GMIM TSSICLRQTD LKSLIAYSSI SHMALVVTAI LIQTPWSFTG

321 LLILCNLAHI LSSSFMFIVI GYMYDNYGVR IFLLLISFF- GISIWSSLFL CLFLFNIDFP FMLLFYVDIF ILYGLISISF
      **    ** * **   *      *     * *  *              *    * *   *             *    * *
    AVIL-MIAHG LTSSLLFCLA NSNYERTHSR IMILSQGLQT LLPLMAFWWL LASLANLALP PTINLLGELS VLVTTFSWSN

401 IYIISFYIIT LTIFLSSIYI YMCLSFYSFV WLDKYLRLDV SINDIYVFMS ISISTIVFYY FIYLLI----    T.brucei URF8
    *          *   * * *         *                   **   *
    ITLLLTGLNM LVTALYSLYM FTTTQWGSLT HHINNMKPSF TRENTLMFMH LSPILLLSLN PDIITGFSSC    Human URF4
```

```
    T.brucei URF10

  1 -----DMFLIF FLFFIMFGFI SGSFMFGR-- NFLSFWLSLV MIIFIVLCMI FSFLMVSVCL YGYYYYDFCL ILMLDFCFIW
                 *                          * *    **      *   *
    MTMHTTMTTL TLTSLIPPIL TTLVNPNKKN SYPHYVKSIV ASTFIIS--L FPTTMFMCLD QEVIISNWHW ATTQTTQLSL

 81 LTYVCSGFYM FIMLLINMVF CFIVFYAFYY MYFDMLLGRF LIIFWIFVVC MNLFILSYDF LTAYCGWELL GLFSFFLISY
             *  **     *    *      * *  *     *      **     *               ***   *  ** ***
    SFKLDYFSMM FIPVAL-FVT WSIMEFSLWY MNSDPNINQF FKYLLIFLIT MLILVTANNL FQLFIGWEGV GIMSFLLISW

161 FWYRFFALKF GFKAFFIGKI GDVLLIFAFS IIFL-SNGFC MTTFYFLNFF CMDYYYIEFS ICLLVGCAFT KSTQFGLHIW
       *  *      *     * **   * *     * **        **                 *   *    ** * *** *
    WYARADANTA AIQAILYNRI GDIGFILALA WFILHSNSWD PQQMALLN-- ---ANPSLTP LLGLLLAAAG KSAQLGLHPW

241 LPDAMEGPIP VSALIHAATL VVCGIILLSF VYWCFDFWFS YFYNLIGWST LILILMT--- --LCVFYNFD VKRYVAFSTI
    ** ***** * **** *  *  ** ** **                 **   * * *   *       *     *  *  *****
    LPSAMEGPTP VSALLHSSTM VVAGIFLLIR FHPLAENSP- ----LIQTLT LCLGAITTLF AAVCALTQND IKKIVAFSTS

321 CQISFSMFCC LCIDIYIGSL FFCYHMFYKA TLFIVLGIWI HIFFGLQDLR CYFFMYFCGC VLARLLLIFA ILNSCSIWFL
     *    *            *   * * * **  **   *  * *     ** *                 * *   *      **
    SQLGLMMVTI GINQPHLAFL HICTHAFFKA MLFMCSGSII HNLNNEQDIR KMGGLLKTMP LTSTSLTI-G SLALAGMPFL

401 CGFYCKDMLL ALLMLLSFYN IIEFLFISII FIFFTMIYNY FLLFFLMFVP KCFCLVDCLF LLFDYECCLV YCLIS----L
     *** **          *  *        **  *       *       *                     *           *
    TGFYSKDHII E-TANMSYTN AWAL---SIT LIATSLTSAY STRMILLTLT GQPRFPTLTN INENNPTLLN PIKGLAAGSL

481 YMCILSIFFI IDFVCIFVFS SYCVFWSFFL NFYNFFDIAI FVVFLILSVG FLYYGCLFFY FFNIDCIMLF WRIFFVIIIL
        *    *                              *      *          *  ** * *
    FAGFLITNNI SPASPFQTTI PLYLKLTALA VTFLGLLTAL DLNYLTNKLK MKSPLC-TFY FSNMLGFYPS ITHRTIPYLG

561 VVFMIFCCWY FVCMIIFMLL FVWNFVIYFR YNLKYCLFFC ILWILYV--- ---------- --    T.brucei URF10
                        *                          **
    LLTSQNLPLL LLDLTWLEKL LPKTISQHQI STSIITSTQK GMIKLYFLSF FFPLILTLLL IT    Human URF5
```

B

T.brucei URF4

```
  1 TPKKSLFLYI IHIFIFLIIY SFIILCDYTT LTLLSFDLLW LLIINLFWIT LLDSYICFIF ILLFLFCFTL FFCFLSFDTR
 81 FLFIIIIIQY IIIFLFIFIN HIIIISILFE IFSLLLFLLL MSSRFGYKIL VLWYYYMLNL INFILLFILL YFMILNYCFF
161 LCDFCFLVFD EEWLGILCLF YTLLILFKLY IAFLILFMEQ LYIRLGVFIF IYMLTFYILF CFILIILLIS FIYFYILFIK
241 LLLFQSCTCV LIGLNSFAIV SLLFVLSVNN FCFLFLIFIS TKNYIFYLYL NFHLIYSISL VLLIIIYYFF IIYNIFDFKY
321 NENYFLINFI FFSFFNNFLI SLLLACLFLC IGAIPIVFGF FIKVFCLLLQ LSYLCICIGF FFIIWLIIIY IFYFRLIVNI
401 FIFSYQFLGF WVVKLSFINI KNLLFFICSS VYILFFDIIN LFDLIL
```

T.brucei URF6

```
  1 DWLWLSRVFD LLCIRTYDFI LWWFDLDFIL YDFVFDFVVC ITFIFIFVLG FFIRIFFSFV FVLLFITFFG ICSLTMLFTG
 81 YYIYYIYILY NFICFFFAFG INFLIYYIEF FIFITFHIFF DFISFSNYIY NYFGILYMFN VMFCAYLFCL FYFVIYFLFC
161 FIFFVIRCLF IVIMDFLFFN FDIFVSILLC DIVYLDFISL LLLYFNFIFN FIYGFFSFVI ILGLLFLLLF LVINLFFGFT
241 FLVYGIQIIL LYYVYWLYMI YSRSCYILMP AILIFFKFIY FDVFFVFVFI LILFIISFFS FFLKDFLFLS LYFDIFGSLY
321 NYDILSYSIF YYQNNQFCLT QLLSIYI
```

T.brucei URF7

```
  1 QLKDNKFQSF YEKHLSTQDI EKTFYKTKIK TKLFLRNYKP KFHQPPLPLL PTKPHTFSPT KTPQSNSPSI SSSSKPNFPK
 81 SPFLSPTNPP FLIDQSQPNL KPLFNKYTKI TINYKKEIYR KHKNKIKLE
```

T.brucei URF9

```
  1 PYKHTDTIKK LNIIHNTINI IITKDNQKLR KFLPNINEPD INPNIIKNKK NIKNISYSKI SSKRCFFNQN KGFPPPKSFW
 81 NSPPRLPFIQ KKLSVGSGWA RTALSLQSHS KNPRLFTLVC YQKCIRDFKR REQSGGRGWE ESRCEQEWSV GETIGLGWMG
161 DGFWGEKFWW GGITEGGKEI CLLYYTYYIN KYIK
```

Figure 5
A) Sequence comparison between *T. brucei* URF5 and human URF1, *T. brucei* URF8 and human URF4 and *T. brucei* URF10 and human URF5. The human sequences are from ref 18. The first methionine is underlined in the *T. brucei* sequences. For further details, see legend to Fig. 3.
B) Amino acid sequence of *T. brucei* URF4, 6, 7 and 9. The first methionine is underlined.

outside the laboratory and only capable of growing in rich culture media.

Sequence analysis of maxi-circles from recently isolated *T. brucei* stocks should shed further light on this possibility.

3) The mitochondrial translational machinery is capable of a -1 frame-shift with a frequency high enough for the production of sufficient cytochrome *c* oxidase subunit II. Such frame-shifts have been postulated to occur as an essential step in the synthesis of proteins produced in low amounts, such as the lysis-protein in the coli-phage MS2 (37) or to explain the leaky phenotype of a yeast mutant (38). In the latter case the frameshift was tentatively attributed to the unique structure of the yeast mitochondrial $tRNA^{phe}$. As yet, we have not identified any mitochondrial tRNA genes and we do not know how such a mechanism could operate in *T. brucei* mitochondria without severely affecting the translation of continuous genes.

4) The mRNA for coxII contains a continuous reading frame as a result of a small splice in the appropriate area of precursor RNA. This

may be a real possibility, in spite of the lack of splices in other T. brucei mitochondrial genes and the lack of precedence for such small splices in (mt)mRNA. Two major RNA-species map in the coxII area (10). Efforts to directly obtain the RNA sequence with the aid of synthetic primers have been hampered so far by the low concentration of these mRNAs in total RNA preparations from T. brucei and attempts to enrich these RNAs by isolation of mitochondria have resulted in extensive degradation. We are currently screening a cDNA clone bank in order to obtain and sequence a coxII-derived cDNA.

The unassigned reading frames

The aminoacid sequence of the unassigned reading frames is given in Fig. 5. Three of them show a low, but distinct homology (20% at the aminoacid level) with mammalian mitochondrial URFs: T. brucei URF5, 8 and 10, and mammalian URF1, 4 and 5 respectively (Fig. 5A). This again follows the pattern that trypanosomal mitochondrial genes are less well conserved, than what is usually encountered (compare e.g. the homology between human and insect URF5: 32%; 31). However, the trypanosomal and corresponding mammalian URFs are of virtually identical size and the hydrophobicity profiles are strikingly similar (not shown), albeit that the T. brucei URFs are slightly more hydrophobic. It is very likely, therefore, that we are dealing with analogous URFs. The first AUG-codon in trypanosomal URFs 4, 5 and 6 occurs rather late in the sequence (around codon 120-140, see Fig. 5 A,B and Table 1). Although this could simply indicate that the proteins start at these positions, sequence alignment of trypanosomal URF5 and human URF1 reveals that the major part of the homologous aminoacids is found upstream of position 148, at which the first AUG codon occurs (Fig. 5A). We have checked the sequence of the 5' half of this gene in a fashion similar to that described for the coxII gene: the same sequence was found in a large number of M13 clones derived both from bloodstream form and cultured T. brucei 427. Therefore, explanations such as sequencing errors, cloning artifacts and/or this gene being a pseudo-gene are unlikely. Although AUA and AUU apparently do not code for methionine (see Table 2), some other unusual codon usage can be envisaged. Close inspection of the 5' sequences of URF4, 5 and 6 reveals the presence of an UUG codon at position 6, 5 and 5, respectively. In URF5 this is, in fact, the only UUG codon of the gene. The possible use of UUG as initiator triplet has been reported for prokaryotes (39). A similar phenomenon may occur in trypanosome mitochondria.

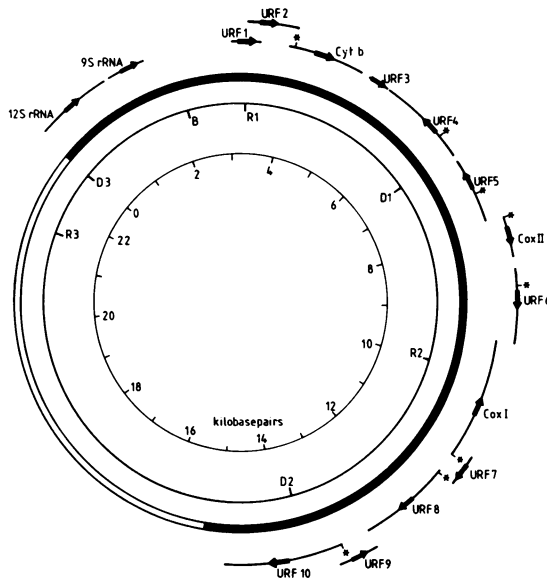

Figure 6
Partial gene map of T. brucei maxi-circle DNA. The arrangement of genes is derived from various studies (6,7,10 and the present data). Arrows indicate direction of transcription of each gene. The black bar indicates the sequenced area (B = BamHI). * indicates the position of the first AUG triplet in each gene.

The URFs discussed sofar display the classical aminoacid composition of mitochondrial membrane proteins, containing a high percentage of hydrophobic residues (70-80%). URF7 and URF9, however, are rich in polar and basic residues (around 55%). These URFs do not contain AUG codons and UUG triplets are found only towards the 3' end of the sequence of URF9. In 2 previous papers (7,10) other examples of such URFs have been discussed (URF1, 2 and 3). At present, we cannot rule out a protein encoding function for those URFs, since abundant transcripts are mapped in some of the URF areas. However, if the unusual properties of the putative URF-proteins are taken into consideration, it seems more likely that these RNAs have some other, as yet unknown function. In this view, the occurrence of reading frames of this length (e.g. URF9 = 194 aminoacids) would be the consequence of constraints imposed on the DNA sequence by the unknown role of the RNA and not by the protein-encoding function.

Gene organization

Figure 6 gives the current state of affairs in the analysis of the gene organization of the T. brucei maxi-circle, based on sequence determination of 15.5 kb (6,7,10, this paper). Comparison with mammalian (18) and insect (31) mitochondrial genomes, which are of similar size and display, between them, a few conserved features (e.g. the order of the rRNA genes and that of the protein genes), reveals a rather unique gene organization in trypanosomes. The order of transcription of the two rRNA genes is reversed: 5'-large rRNA-small rRNA-3', the order of the protein genes is different, coxI being transcribed in a direction opposite to that of the coxII and apocytochrome b genes and tRNA genes are conspicuously absent. In fact, the only point of similarity is the relative position of URF8 and 10: URF4 and 5 in mammalian and insect DNA, which are homologous to these trypanosomal URFs, also occur in tandem in their respective DNAs.

tRNA genes are used as processing points in the expression of mammalian and (possibly) insect mitochondrial genes. Their absence in the sequences obtained sofar may be explained in a few alternative ways:

a) tRNAs in trypanosome mitochondria have a highly unusual structure, which allows them to go undetected in a computer analysis looking for classical or semi-classical tRNAs.

b) The genes are clustered in the still unsequenced part of the maxi-circle, or

c) tRNAs are imported, as has been postulated for some of the Tetrahymena mitochondrial tRNAs (40) or are encoded on mini-circles.

We are currently completing the maxi-circle sequence of T. brucei and comparing relevant areas to maxi-circle sequences of the insect trypanosome, Crithidia fasciculata, in order to gain more insight in this intriguing problem. This approach will also be followed to localize the remaining protein genes and URFs and to find out whether some of the unusual features discussed above (the high cysteine content of proteins, the discontinuous coxII gene, the possible use of UUG as initiator codon) are shared by other trypanosomatidae.

ACKNOWLEDGEMENTS

We thank Prof.Dr.P.Borst, Prof.Dr.L.A.Grivell and Dr.H.F.Tabak for their continuing interest, stimulating discussions and critical reading of the manuscript. We are indebted to Dr.A.O.Muijsers for his comments on the coxI and coxII sequences, to Ms G.J.M.Scholts for expert typing, to Mrs F.Fase-Fowler for a kind gift of cultured T. brucei 427 and to

E.Tigchelaar, K.Gravenmeijer and Ms J.van den Burg for skilful technical assistance. We thank J.Bruning for sharing with us his unpublished results on the nucleotide sequence of a cDNA derived from a coxI transcript. This work was supported in part by a grant to L.A.Grivell from the Foundation for Fundamental Chemical Research (SON), which is subsidized by The Netherlands Organization for the Advancement of Pure Research (ZWO).

Abbreviations: kb, kilo base pairs; URF, unassigned reading frame; coxI(II), cytochrome c oxidase subunit I(II); mt, mitochondrial.

+Present address: Istituto di Anatomia Comparata Biologia Generale, Genetica, Via Romana 17, Firenze, Italy

§To whom correspondence should be addressed

REFERENCES

1 Simpson, L. (1972) Intern.Rev.Cytol. 32, 139-207.
2 Borst, P. and Hoeijmakers, J.H.J. (1979) Plasmid 2, 20-40.
3 Englund, P.T. (1981) in: Biochemistry and Physiology of Protozoa (Levandowsky, M. and Hutner, S.H., Eds), 2nd edn, Vol.4, Academic Press, New York, pp. 334-383.
4 Stuart, K. (1983) Mol.Biochem.Parasitol. 9, 93-104.
5 Hoeijmakers, J.H.J., Snijders, A., Janssen, J.W.G. and Borst, P. (1981) Plasmid 5, 329-350.
6 Eperon, I.C., Janssen, J.W.G., Hoeijmakers, J.H.J. and Borst, P. (1983) Nucl.Acids Res. 11, 105-125.
7 Benne, R., De Vries, B.F., Van den Burg, J. and Klaver, B. (1983) Nucl.Acids Res. 11, 6925-6941.
8 Johnson, B.J.B., Hill, G.C., Fox, T.D. and Stuart, K. (1982) Mol. Biochem.Parasitol. 5, 381-390.
9 Simpson, L., Spithill, T.W. and Simpson, A.M. (1982) Mol.Biochem. Parasitol. 6, 253-264.
10 Benne, R., Agostinelli, M., De Vries, B.F., Van den Burg, J., Klaver, B. and Borst, P. (1983) in: Mitochondria 1983: Nucleo-Mitochondrial Interactions (Schweyen, R.J., Wolf, K. and Kaudewitz, F., Eds), De Gruyter, Berlin, pp. 285-302.
11 Woese, C.R., Gutell, R., Gupta, R. and Noller, H.F. (1983) Microbiol.Rev. 47, 621-669.
12 Noller, H.F., Kop, J.A., Wheaton, V.W., Brosius, J., Gutell, R.R., Kopylov, A.M., Dohme, F. and Winship, H. (1981) Nucl.Acids Res. 9, 6167-6189.
13 Birnboim, H.C. and Doly, J. (1979) Nucl.Acids Res. 7, 1513-1523.
14 Borst, P., Fase-Fowler, F., Hoeijmakers, J.H.J. and Frasch, A.C.C. (1980) Biochim.Biophys.Acta 610, 197-210.
15 Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E. and Surrey, S. (1982) Proc.Natl.Acad.Sci.U.S. 79, 4298-4302.
16 Van der Marel, G.A., Marugg, J.E., De Vroom, E., Wille, G., Tromp, M., Van Boeckel, C.A.A. and Van Boom, J.H. (1982) Recl.Trav.Chim. Pays-Bas 101, 234-241.
17 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc.Natl.Acad. Sci.U.S. 74, 5463-5467.
18 Anderson, S., Bankier, A.T., Barrell, B.G., De Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A.,

Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R. and Young, I.G. (1981) Nature 290, 457-465.
19 Bonitz, S.G., Coruzzi, G., Thalenfeld, B.E., Tzagoloff, A. and Macino, G. (1980) J.Biol.Chem. 255, 11927-11941.
20 Coruzzi, G. and Tzagoloff, A. (1979) J.Biol.Chem. 254, 9324-9330.
21 Coruzzi, G., Bonitz, S.G., Thalenfeld, B.E. and Tzagoloff, A. (1981) J.Biol.Chem. 256, 12780-12787.
22 Fox, T.D. (1979) Proc.Natl.Acad.Sci.U.S. 76, 6534-6548.
23 Macino, G. and Tzagoloff, A. (1980) Cell 20, 507-517.
24 Nobrega, F.G. and Tzagoloff, A. (1980) J.Biol.Chem. 255, 9828-9837.
25 Thalenfeld, B.E. and Tzagoloff, A. (1980) J.Biol.Chem. 255, 6173-6180.
26 Tzagoloff, A., Nobrega, M., Akai, A. and Macino, G. (1980) Current Genet. 2, 149-157.
27 Capaldi, R.A., Malatesta, F. and Darley-Usmar, V.M. (1983) Biochem. Biophys.Acta 726, 135-148.
28 Steffens, G.J. and Buse, G. (1979) Hoppe-Seyler's Z.Physiol.Chemie 360, 613-619.
29 Welinder, K.G. and Mikkelsen, L. (1983) FEBS Lett. 157, 233-239.
30 Grivell, L.A. (1983) Sc.Amer. 248, 78-89.
31 Clary, D.O., Wahleithner, J.A. and Wolstenholme, D.R. (1984) Nucl. Acids Res. 12, 3747-3762.
32 Opperdoes, F.R. (1977) Ph.D.Thesis, Amsterdam.
33 Hoeijmakers, J.H.J. and Weijers, P.J. (1980) Plasmid 4, 97-116.
34 Ray, S.K. and Cross, G.A.M. (1972) Nature New Biol. 237, 174-175.
35 Hill, G.C. (1972) in: Comparative Biochemistry of Parasites (Van den Bossche, H., Ed), Acad.Press, New York and London, pp. 395-415.
36 Evans, D.A. and Brown, R.C. (1972) J.of Protozool. 19, 365-369.
37 Kastelein, R.A., Remaut, E., Fiers, W. and Van Duin, J. (1982) Nature 295, 35-41.
38 Fox, T.D. and Weiss-Brummer, B. (1980) Nature 288, 60-63.
39 Young, I.G., Rogers, B.L., Campbell, H.D., Jaworowski, A. and Shaw, D.C. (1981) Eur.J.Biochem. 116, 165-170.
40 Suyama, Y. and Hamada, J. (1976) in: Genetics and Biogenesis of Chloroplasts and Mitochondria (Bücher, Th., Neupert, W., Sebald, N. and Warner, S., Eds), North-Holland, Amsterdam, pp. 763-770.