Achievements and Perspectives of Mitochondrial Research
Volume II: Biogenesis, E. Quagliariello et al. editors

GENOMIC ORGANIZATION AND TRANSCRIPTION OF MITOCHONDRIAL MAXICIRCLE DNA IN TRYPANOSOMID PROTOZOA

L. SIMPSON, A. M. SIMPSON, V. DE LA CRUZ, N. NECKELMANN AND M. MUHICH
Biology Department and Molecular Biology Institute, University of California, Los Angeles, CA 90024 (USA)

INTRODUCTION

Comparative analysis of mitochondrial genomes from diverse organisms may lead to a deeper understanding of the mode of evolution of this originally endosymbiotic bacterial genome into the present day organelle genomes of limited but fairly stable genetic content. Such analysis may in addition yield an understanding of the functional significance of organellar genomic segregation for the life of the present day eukaryotic cell. Animal and fungal mitochondrial systems have been most intensively studied, with very few representatives of other lower eukaryotic cell types.

The trypanosomids or kinetoplastid protozoa represent a large and somewhat diverse class of lower eukaryotic cells that are characterized by the existence of a single mitochondrion per cell which has a unique type of mitochondrial DNA known as kinetoplast DNA (1, 2, 3). This DNA consists of a single giant network of thousands of catenated minicircles and 20-50 catenated maxicircles. The function of the minicircle DNA is still a mystery; most evidence indicates a lack of genetic activity (4), although there is recent preliminary evidence for a possible minicircle coding function in Crithidia fasciculata (5). The maxicircle has been shown by sequence analysis to represent the informational DNA species homologous to that in other organisms. Substantial maxicircle sequence data is now available for two species, Leishmania tarentolae (6, 7, 8) and Trypanosoma brucei (9, 10, 11, 12, 13, 14) and limited sequence data is available for Crithidia fasciculata (15), Leptomonas (unpublished results), and Trypanosoma cruzi and Herpetomonas samuelpessoa (Kidane and Morel, personal communication). The cross-species sequence comparisons are interesting not only in terms of the changes in mitochondrial genomic sequences and organization between trypanosomids and humans and yeast, but also in terms of the phylogeny of these different species within the family, Trypanosomatidae.

In this paper we will discuss results from our laboratory on the genomic organization and transcription of the maxicircle genome of L. tarentolae, with emphasis on comparative aspects of the data.

RESULTS AND DISCUSSION

Genetic coding potential of L. tarentolae maxicircle DNA

The sequence of a continuous stretch of 21 kb of the approximately 30 kb maxicircle of L. tarentolae is known (6, 16) (unpublished results). This includes the entire 16.7 kb transcribed region and 4.3 kb of the untranscribed "divergent region". The 9S and 12S kinetoplast ribosomal RNA genes and seven structural genes were identified by comparison of the DNA and translated amino acid sequences with those of known genes from other organisms. The structural genes- COI, COII, COII, CYB, ND4 (HURF4), ND5 (HURF5) and ND1 (HURF1)- were identified by two basic methods- analysis of the hydropathy patterns (17) of the open reading frames (ORFs), and determination of the statistical significance of amino acid alignments with known genes by the SEQDP computer program of Kanehisa (18) (Fig. 1).

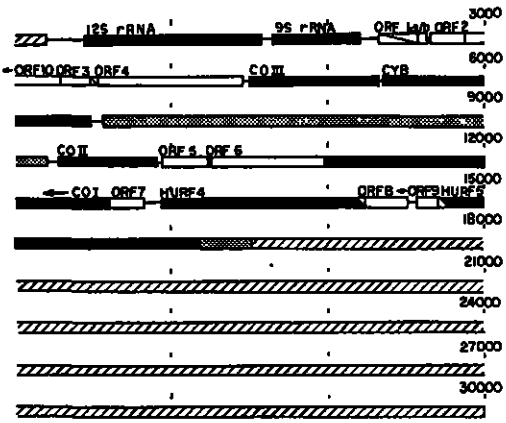

Fig. 1. Genomic organization of the maxicircle DNA of L. tarentolae. The non-transcribed divergent region is indicated by cross-hatching, and the identified genes are indicated by dark shading. Unidentified open reading frames (ORFs) are blank. All identifed genes except for COI (and ND1) are transcribed left to right (arrow). The portions of the transcribed region the sequence of which is not published are indicated by stippling. Numbers at right indicate base pairs from the EcoRI site. HURF4 and HURF5 = ND4 and ND5. From de la Cruz et al (6) by permission.

The sequence of the region between the CYB gene and the COII gene was recently obtained (unpublished results), and the ND1 gene identified as well as several ORFs (ORF11, ORF12, ORF13). Genes encoding subunits of the mitochondrial F1 ATPase have not been identified by these methods. It is possible that the ATPase subunit 6, 8 and 9 genes are nuclear-localized in this species, but it is also possible that the mitochondrial genes have diverged extensively and are not

easily recognizable. One possible candidate for ATPase 8 and 6 is the ORF3-4 region, which has a single transcript covering both ORFs (19) and has a hydropathic pattern with some similarity to that of human or yeast ATPase 8 and 6; the SEQDP alignment values, however, are not statistically significant.

Maxicircle genes are greatly diverged from homologous genes in yeast and animal mitochondria

The extent of nt identity between the putative maxicircle genes and known genes varies from 38 to 57% and the extent of amino acid identity varies from 18 to 42% (6). However the hydropathic patterns of the maxicircle genes are strongly conserved (Figure 2). The statistical significances of the computer-derived alignments of the amino acid sequences of the maxicircle genes and known genes are shown in Table 1. Clearly, the COI, COII, CYB, ND4 and ND5 (and ND1) alignments are quite significant. The alignment of the maxicircle putative COIII sequence, however, is barely significant statistically, but there is reason to believe that this maxicircle gene does represent a highly diverged homologue of the COIII gene in humans and yeast (6). In all cases, however, the statistical "distance" of the maxicircle genes from both human and yeast is greater than the distance of human and yeast genes from each other, implying that the maxicircle genes have diverged more from the homologous human and yeast genes than the human and yeast genes have diverged from each other. The extent of sequence similarity varies from the highly conserved COI, COII and ND5 genes to the moderately conserved CYB and ND1 and ND4 genes, to the poorly conserved COIII gene.

One interpretation of the high degree of divergence of the maxicircle gene sequences from the homologous fungal and animal sequences is to postulate an extremely early divergence of the kinetoplastid protozoal cell line in the eukaryotic cell lineage. This hypothesis is consistent with a phylogenetic analysis from cytochrome c data of Crithidia (20) and an analysis of the sequence of the 5.8 S rRNA of T. brucei and C. fasciculata (21, 22).

Properties of maxicircle genes

There is a pronounced strand asymmetry in all the identified maxicircle genes and in several ORFs, with T being the predominant base in the sense strand; the T/A ratio is approximately 2 (6). Assuming that the T/A strand asymmetry is a characteristic of a maxicircle gene, we can predict that ORF4, ORF5, ORF6 and possibly ORF7 and ORF8 represent genes of unknown function. Analysis of codon bias represents another method to identify coding regions. Application of the Pustell (23) codon bias computer program to the L. tarentolae maxicircle sequence, using the identified genes to establish the codon bias table, indicates a possible coding function for ORFs 10, 3, 4, 5, 6, 11 and 12 (unpublished results).

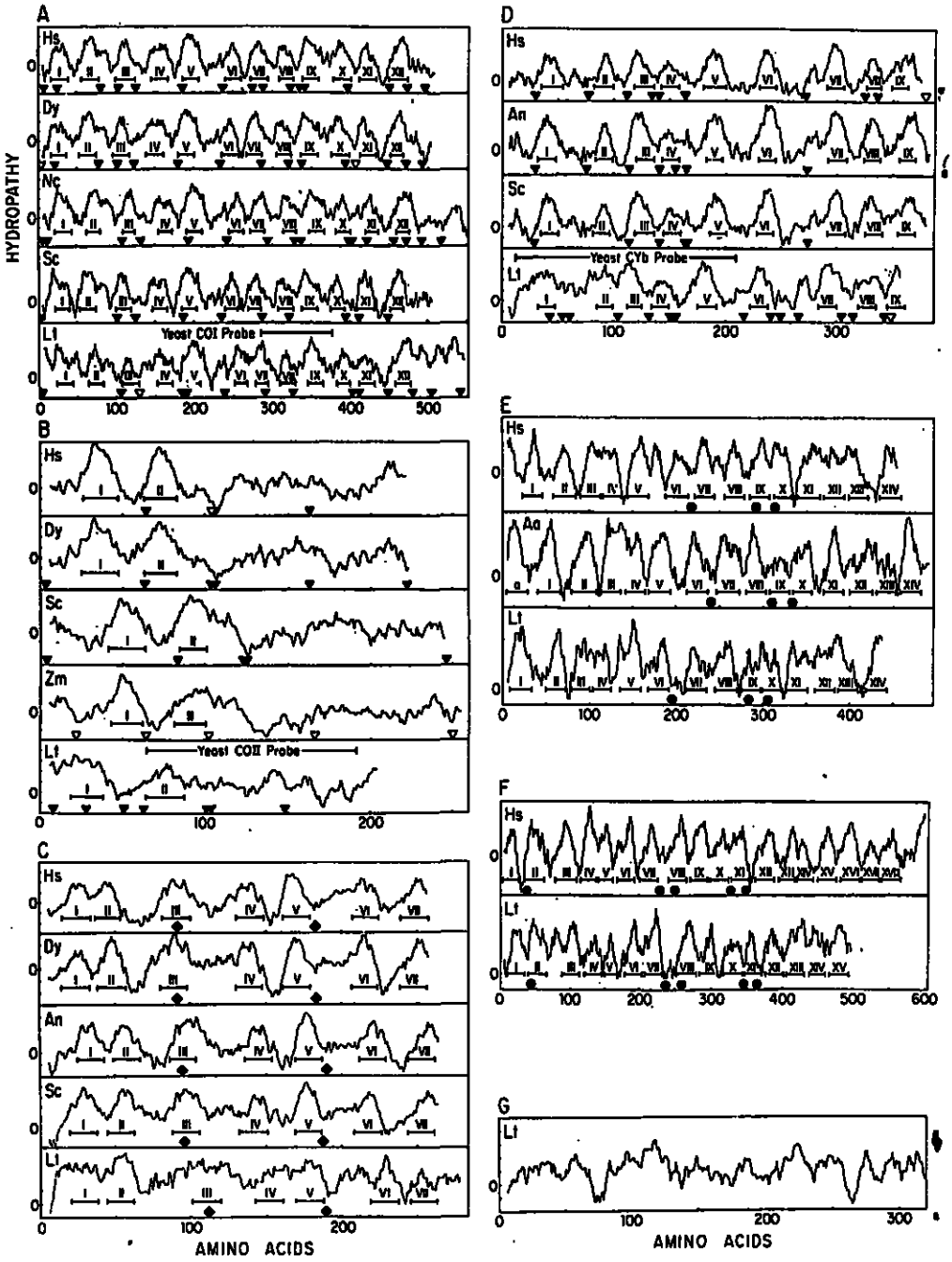

Fig. 2. Hydropathy patterns of selected known mitochondrial proteins and the putative L. tarentolae maxicircle proteins. Hydropathy patterns were obtained

from the amino acid sequences, using a window of 11 amino acids. The tic marks on the y axis are in units of 10; hydrophobic values are positive (up) and hydrophilic values are negative (down). The horizontal bars indicate hydrophobic domains which are presumed to be membrane-spanning regions within the proteins; the bars are numbered to facilitate comparisons with homologous domains in related proteins. In plots A, B, and D, TGA tryptophan residues are indicated by solid inverted triangles and TGG tryptophan residues by open inverted triangles. In plot C, conserved glutamic acid residues are indicated by solid circles. Regions of homology with heterologous yeast petite hybridization probes are indicated for COI, COII, and cytochrome b of L. tarentolae. Abbreviations Hs, human; Dy, Drosophila yakuba, Nc, Neurospora crassa; Sc, Saccharomyces cerevisiae; Zm, Zea mays; Aa, Aspergillus amstelodemi; An, Aspergillus nidulans; Lt, L. tarentolae. A, COI proteins; B, COII proteins; C, COIII proteins; D, cytochrome b proteins; E, URF4 proteins; F, URF5 proteins; G, ORF4 from the L. tarentolae maxicircle. From de la Cruz et al. (6) by permission.

The trypanosome mitochondrial genetic code

From an analysis of codon usage in the maxicircle COI and CYB genes, we were able to conclude that the only apparent variation from the universal genetic code is the assignment of TGA as tryptophan (6). The ATA codon is probably isoleucine and may be an initiator of protein synthesis. The possibility of noncanonical initiation codons exists since several ORFs do not have a satisfactory start codon. Although there are more cysteines in the maxicircle genes than in the homologous genes from other organisms, there is no consistent pattern of amino acid replacements by cysteine and therefore no evidence for the reassignment of TGT or TGC to different amino acids. The presence of multiple cysteine residues in a protein could have significant effects on its conformation and it should prove of interest to obtain amino acid sequences of purified maxicircle gene products to verify this assignment.

Several genes have apparent reading frame shifts

There is a shift in reading frame between amino acids 168 and 172 in the COII sequence (6). This frameshift is not due to a DNA sequencing error, or to errors introduced in the cloning process since the identical sequence was obtained using native (uncloned) maxicircle DNA (unpublished results). A similar frameshift in an equivalent position occurs in the published sequences of the COII gene from two strains of T. brucei (9, 12). A similar situation seems to exist for the ORF5-6 region; these overlapping ORFs are covered by a single 1000 nt transcript (19). However, in the case of T. brucei this region is present as a single ORF (ORF2A) which shows good amino acid sequence similarity with the translated ORF5-6 sequence (unpublished results).

There are several possibilities to explain the frameshifts in at least two maxicircle genes. One is that the majority of maxicircle molecules contain pseudogenes but there is a small population of maxicircle molecules that contain the transcribed genes lacking frameshifts. The difficulty with this hypothesis

TABLE I

SIGNIFICANCE OF SEQUENCE SIMILARITY BETWEEN PAIRS OF PROTEINS

SEQDP program of Goad and Kanehisa. The values represent the average number of standard deviation units for 10 SEQDP trials (10 x 10 random sequence comparisons), indicating the significance that the calculated distance between two sequences is not due to chance.

Standards versus standards

Gene	H/D[a]	H/Y	H/A	H/N	H/Z	H/T	Y/A	Y/N	Y/Z	Y/T	A/T
COI	116±44[b]	93±23		94±17				100±18			
COII	56±12	45±14			41±7				58±12		
COIII	66±12	51±16	53±18	58±12			53±12	55±17			
CYb		61±14	65±18			21±6				15±2	19±5
HURF4			29±8								

Standards versus L. tarentolae putative genes

Gene	H	D	Y	A	N	Z	T
COI	62±20	72±22	65±18		54±15		
COII	21±5	22±8	18±5			15±6	
COIII	5±2	6±1	4±1	7±2	8±2		
CYb	17±3		18±6	22±3			75±24
HURF4	10±4			10±3			
HURF5[c]	21±5						

[a] Abbreviations H, Human; D, Drosophila; Y, Saccharomyces cerevisiae; A, Aspergillus; N, Neurospora; Z, Zea mays; T, T. brucei; CYb, cytochrome b.

[b] Mean number of S.D. units ± S.D. (10 trials).

[c] Only the first 510 amino acids of HURF5 were used in this comparison because the complete sequence was not available. From de la Cruz et al. (6) by permission.

is that extensive cloning of the COII gene from both T. brucei (9, 12) and L. tarentolae maxicircle DNA was performed and the identical frameshift was present in all isolates. The cross-species conservation of the COII frameshift also argues against a pseudogene explanation. Other possible explanations proposed by Hensgens et al (9) include the existence of a translational -1 frameshift or a small splice in the precursor RNA. Further work must be done to clarify this problem.

Transcription of maxicircle genes

Stable transcripts for all identified maxicircle structural genes (except ND1), most of the ORFs, and the 9S and 12S rRNA genes were mapped (Fig. 3), and the locations of the 5' ends determined by primer runoff using synthetic oligonucleotides (19). All genes except ND1 and COI (and ORF11) are transcribed from the same strand as the 12S and 9S RNA genes. In the cases of ORF3-4 and ORF5-6, a single transcript covers two contiguous overlapping reading frames, implying that these may represent single genes as discussed above. The 5' ends of the stable RNAs for all genes analyzed are located 20-64 nt from the putative translation initiation codons. It is not known which if any of the stable mRNAs represent primary unprocessed transcripts. The only consistent promotor-like consensus sequence is a short AT-rich region immediately 5' adjacent to the presumed mRNA initiation sites. An attempt to identify primary transcripts by specifically labeling 5' di- or triphosphate ends of total kinetoplast RNAs with guanylyltransferase yielded two major labeled species that comigrated with 9S and 12S RNAs and at least four higher molecular weight species of 1.2, 1.5, 1.9 and 4 kb (19). These preliminary results suggest the existence of separate promotors for the 12S and 9S RNAs and possibly several additional promotors for mRNAs, but this problem requires further work.

The stable transcripts of the maxicircle genes are present in different abundances; for example, the 1000 nt ORF5-6 transcript is of low abundance compared to the 1.8 kb COI transcript and the 700 nt COII transcript.

Several small transcripts were noted which correspond to regions that do not meet the criteria of coding strand T/A asymmetry and codon bias which are characteristic of the identified maxicircle genes; for example a 320 nt RNA corresponds in location and polarity to ORF1A/B.

The tRNA problem

Several putative tRNA cloverleaf structures can be formed from the DNA sequence at the sites indicated by arrows in Figure 3, upstream of the 12s RNA gene and between ORF2 and ORF3. However due to the extensive variability of tRNAs in other mitochondrial systems, it is impossible to establish the existence of mitochondrial tRNA genes by inspection alone. In any case, the number of putative tRNA genes in the maxicircle sequence is limited and could not comprise a complete set of 22 tRNA genes as found in animal and fungal mitochondria. It is of interest that Suyama (24) has proposed that only 10 of the mitochondrial tRNAs of the ciliated protozoan, Tetrahymena, are coded by mitochondrial DNA and that the remainder are imported from the cytoplasm. We would like to raise the possibility that this may be a general phenomenon among the protozoa, perhaps related to the extremely ancient divergence of these lines

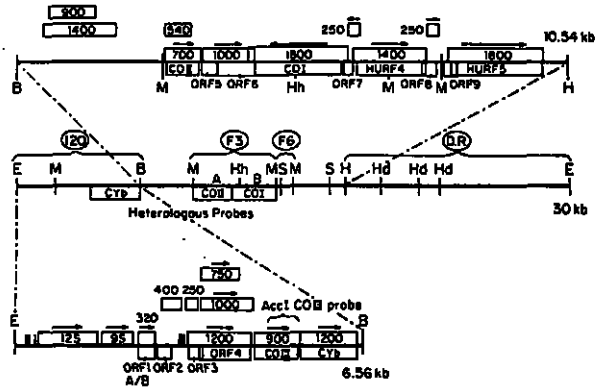

Fig. 3. Partial transcription map of L. tarentolae maxicircle. The restriction map of the entire EcoRI linearized maxicircle is shown in the center with the relevant cloned and sequenced fragments indicated by brackets. The non-transcribed divergent region (DR) is also indicated. Beneath the center map are presented the approximate localizations of the structural genes as determined previously by heterologous hybridizations. The "120 region" is expanded on the bottom and the remaining "F3F6F1" (refers to the EcoRI/BamHI/MspI maxicircle fragments) transcribed region is expanded on the top, with the identified genes and ORFS indicated beneath the lines and the well localized transcripts on the line. Poorly localized transcripts are indicated in some cases by boxes above the line. Only major RNA species are shown on this map since the origins and localizations of the minor species have not yet been determined. The direction of transcription if known is shown by an arrow. The vertical arrows indicate the location of putative tRNA genes. The transcript localizations were derived from the previously published Northern blot analysis using restriction fragment probes and from blot analysis and primer runoff analysis. From Simpson et al. (19) by permission.

from the lines that led to fungal and animal cells. This is obviously a problem that requires further work.

The 9S and 12S maxicircle transcripts represent minimal ribosomal RNAs

Determination of the 5' and 3' ends of the 9S and 12S RNAs and the availability of the sequences from T. brucei allowed the construction of secondary structure models (7, 8) based on those developed for E. coli (25). The 610 nt 9S RNA exhibits a minimal secondary structure in which all four domains of the E. coli 16S rRNA structure are preserved; however, some stems and loops have been greatly reduced or eliminated entirely. The 1173 nt 12S RNA also exhibits a secondary structure which is equivalent in certain respects to the corresponding portions of the E. coli 23S rRNA model. A complete secondary structure for the 12S RNA could not be constructed using only the L. tarentolae and the T. brucei sequences for verification of helical regions, but a conservation of several

regions known to have functional significance was noted. The conserved regions include the alpha sarcin cleavage site, the puromycin binding locus, the peptidyl transferase center and several regions of unknown function known to interact with the peptidyl transferase center. Several large regions are not present in the 12S RNA model, suggesting that these regions are not crucial for translation. These results indicate that the 12S and 9S kinetoplast RNAs represent unusually small, highly diverged mitochondrial ribosomal RNAs, the study of which may prove instructive in indicating regions of rRNA that are absolutely crucial for basic translational functions.

The nontranscribed divergent region of the maxicircle contains a diverse set of repetitive sequences

The L. tarentolae maxicircle genome consists of a 16.7 kb conserved transcribed region that contains rRNA and structural genes and a 13 kb non-transcribed region that shows a lack of sequence similarity between species (26). We have termed this the "divergent" region since length and sequence variation in this region is largely responsible for the size differences observed between maxicircles of different species and genera. A 4.3 kb segment of the divergent region extending from the end of the ND5 gene was sequenced (16) (unpublished results). The segment consisted almost entirely of head-to-tail repeated sequences which can be grouped into at least six families. The three simplest repeats (A, B, E) are clustered into large arrays (Table II, Fig. 4). The existence of rapidly evolving repetitive sequences within the divergent region could account for the large variation in size as well as the lack of cross-species sequence homology in this region of the maxicircle. Unequal recombinational exchange could represent one possible mechanism for the rapid rate of divergent region sequence change.

Comparison of the maxicircle genomes of L. tarentolae and T. brucei

The entire 16.7 kb transcribed sequence of the L. tarentolae maxicircle was compared to the 15 kb transcribed sequence of the T. brucei maxicircle by Diagon (27) dot matrix analysis (unpublished results). This analysis confirmed our previous hybridization results (26) that the transcribed regions of the two genomes were arranged colinearly with several regions of nonhomology interspersed. The main region of nonhomology is a 3.1 kb sequence between the 9S RNA gene and the CYB gene in L. tarentolae, which is substituted by a 2.1 kb sequence in T. brucei. The L. tarentolae sequence contains the COIII gene and several ORFs. Within this nonhomologous region there are, however, two short sequences which show similarity, the significance of which is unclear. The absence of the COIII gene in the T. brucei maxicircle is striking and raises the questions as to whether this is now a nuclear gene, whether the sequenced molecule is a pseudogene, or whether the cytochrome oxidase enzyme actually lacks subunit III in this species.

TABLE II

CONSENSUS NUCLEOTIDE SEQUENCES AND FREQUENCY OF APPEARANCE OF REPEATING ELEMENTS IN THE Lt30-54 SEGMENT OF THE L. TARENTOLAE MAXICIRCLE DIVERGENT REGION.

Repeat	Consensus Nucleotide Sequence	Reiteration Number
A	AATAATAT	29
B	AAATT	103[c]
C	33-239 nt[a]	10
D	149 nt[b]	2
E	ATATTAAAACAAGTTATTCCC	10
F	ACAAATTTTGACAGAGTTCTATAAA	2
	ATTAGACAAACGTACTTAAAACTGTC	

a See Fig. 5 in ref. 16 for nucleotide sequences
b nt 1-149 and 1877-2026, Fig. 2 in ref. 16
c Number of repeating elements present in clusters of five or more contiguous units. From Muhich et al. (16) by permission.

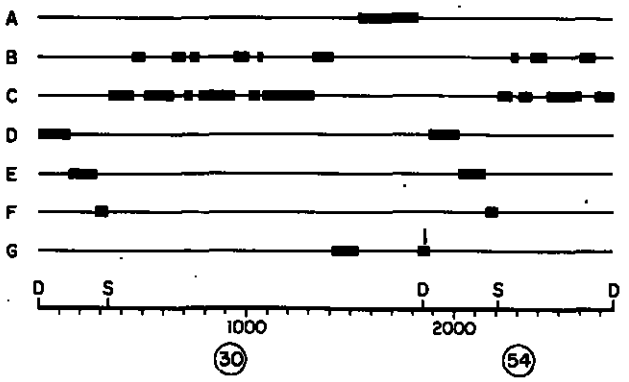

Fig. 4. Diagramatic representation of the distribution of the various types, -A, -B, -C, -D, -E, -F and -G, of repetitive sequences comprising the Lt30 and Lt54 fragments of the L. tarentolae maxicircle divergent region. The boxes in G represent unique sequences in that they are each present only once in the entire 2759 bp. The arrow indicates the two fold axis of symmetry of a 47 nucleotide palindrome. From Muhich et al. (16) by permission.

In the remainder of the analyzed regions, the identified genes and ORFs are conserved both in location and polarity, and the intergenic regions are not

conserved at all except in size. These nonconserved regions include ORF12 (between CYB and ORF11), ORF7 (between COI and ND4), ORFs 8 and 9 (between ND4 and ND5). Absolute amino acid and nucleotide matches of the conserved genes and ORFs range from 71 to 84%, and SEQDP values for the alignments are highly significant and range from 149 to 50 standard deviation units. In all cases there is a preponderance of transversions over transitions.

Conclusions

The maxicircle genome of L. tarentolae is similar to the mitochondrial genomes of animal and fungal cells but has several unique characteristics. It is tightly packed and extensively transcribed but in addition contains a large non-transcribed region that consists mainly of tandemly repeated sequences. Several of the genes contain frameshifts, at least one of which (in the COII gene) is conserved across species barriers. The identified genes consist of two mini-ribosomal RNA genes, three subunits of cytochrome oxidase, cytochrome b, and three genes homologous to human mitochondrial NADH dehydrogenase subunits 1, 4 and 5. ATPase subunit genes have not been identified. Several unidentified open reading frames are also present which have the characteristics of genes. The presence of tRNA genes is uncertain, but it is likely that there is probably not a complete set of 22 tRNA genes as in animal and fungal mitochondria.

A comparison of the maxicircle genomes of L. tarentolae and T. brucei indicates that most identified genes and ORFs are conserved in location and polarity, but that several intergenic regions are not conserved. There is also an absence of the COIII gene and several ORFs from the sequenced T. brucei genome. The existence of a large number of easily cultivatable trypanosomid species from several different genera and the biological evidence for a phylogenetic progression from the more primitive genera such as Crithidia and Leptomonas to the more recently evolved genera such as Leishmania and Trypanosoma provide a unique opportunity to study the mitochondrial phylogeny of a large group of related organisms. Such a comparative study may lead to a deeper understanding of the nature and mode of evolution of organelle genomes in general and also should increase our knowledge of this ancient branch of the eukaryotic tree.

ACKNOWLEDGEMENTS

This work was supported by research grants AI09102 and AI13027 from the NIH. We acknowledge the collaboration of Drs. Ken Stuart and Jean Feagin in the comparative analysis of the T. brucei maxicircle sequence.

REFERENCES

1. Simpson L (1972) Intern Rev Cytol 32:139-207
2. Stuart K (1983) Mol Biochem Parasitol 9:93-104
3. Simpson L (1985) Intern Rev Cytol, In press
4. Kidane G, Hughes D, Simpson L (1984) Gene 27:265-277
5. Schlomai J, Zadok A (1983) Nucl Acids Res 11:4019-4034
6. de la Cruz V, Neckelmann N, Simpson L (1984) J Biol Chem 259:15136-15147
7. de la Cruz VF, Lake JA, Simpson AM, Simpson L (1985) Proc Natl Acad Sci USA 82:1401-1405
8. de la Cruz V, Simpson A, Lake J, Simpson L (1985) Nucl Acids Res, In press
9. Hensgens A, Brakenhoff J, De Vries B, Sloof P, Tromp M, Van Boom J, Benne R (1984) Nucl Acids Res 12:7327-7344
10. Benne R, DeVries B, Van den Burg J, Klaver B (1983) Nucl Acids Res 11:6925-6941
11. Feagin J, Jasmer D, Stuart K (1985) Nucl Acids Res, In press
12. Payne M, Rothwell V, Jasmer D, Feagin J, Stuart K (1985) Mol Biochem Parasitol 15:159-170
13. Feagin J, Stuart K (1985) Proc Natl Acad Sci USA 82:3380-3384
14. Jasmer D, Feagin J, Payne M, Stuart K (1985) Mol Cell Biol, In press
15. Sloof P, Van den Burg J, Voogd A, Benne R, Agostinelli M, Borst P, Gutell R, Noller H (1985) Nucl Acids Res 13:4171-4190
16. Muhich M, Neckelman N, Simpson L (1985) Nucl Acids Res 13:3241-3260
17. Kyte J, Doolittle R (1982) J Mol Biol 157:105-132
18. Kanehisa MI (1982) Nucl Acids Res 10:183-196
19. Simpson AM, Neckelmann N, de la Cruz V, Muhich M, Simpson L (1985) Nucl Acids Res, In press
20. Schwartz RM, Dayhoff MO (1978) Science 199:395-403
21. Hasan G, Turner M, Cordingley J (1984) Cell 37:333-341
22. Dorfman D, Lenardo M, Reddy L, Van der Ploeg L, Donelson J (1985) Nucl Acids Res 13:3533-3549
23. Pustell J, Kafatos FC (1984) Nucl Acids Res 12:643-655
24. Suyama Y (1982) In: Slonimski P, Borst P, Attardi G (eds) Mitochondrial Genes. Cold Spring Harbor Laboratory, New York:449-455
25. Woese C, Gutell R, Gupta R, Noller H (1983) Microbiol Rev 47:621-669
26. Muhich M, Simpson L, Simpson AM (1983) Proc Natl Acad Sci USA 80:4060-4064
27. Staden R (1982) Nucl Acids Res 10:2951-2961