Gene, 141 (1994) 63–69

GENE 07731

The nuclear genomes of African and American trypanosomes are strikingly different

(Codon usage; genome compartmentalization; *Trypanosoma brucei*; *T. cruzi*)

Héctor Musto*, Helena Rodríguez-Maseda* and Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 75005 Paris, France

Received by P.A.M. Michels: 4 October 1993; Revised/Accepted: 3 November/4 November 1993; Received at publishers: 29 November 1993

SUMMARY

We have investigated the compositional distributions of exons and their different codon positions, as well as the codon usage and amino-acid (aa) composition of the nuclear genomes of the African and American trypanosomes *Trypanosoma brucei* and *T. cruzi*. Very large differences between the two species were found in all the properties investigated. The most striking differences concern the compositional distributions of third codon positions and the extremely large nucleotide divergence of third codon position for homologous genes encoding proteins that are highly conserved in their aa sequences. Moreover, if coding sequences from each species are divided into two groups according to the GC levels in third codon positions, very different codon usages and aa compositions are found. This indicates a compositional compartmentalization in both genomes which had previously been detected in *T. brucei* (and *T. equiperdum*) by compositional fractionation.

INTRODUCTION

The species of the genus *Trypanosoma* (Class *Zoomastigophorea*, Order *Kinetoplastida*) are flagellated protozoan parasites of vertebrates (Soulsby, 1982). The most intensively studied species are *T. brucei* and *T. cruzi*, which cause serious diseases in human beings. The *T. brucei* complex is responsible for African sleeping sickness, and *T. cruzi* for Chagas' disease in South America (de Raadt and Seed, 1977; Fife, 1977). In both cases, the parasites are transmitted from one human host to the next by insects.

Apart from sanitary and economical reasons, much of the interest in trypanosome research comes from several remarkable aspects of their biochemistry and molecular biology. Among these are the presence of glycosomes (Opperdoes, 1985), the kinetoplast DNA network (Simpson, 1986), the presence of about 100 minichromosomes for a nuclear haploid genome of only 3.7×10^7 bp (Michels, 1987), the post-transcriptional editing of mitochondrial mRNAs (Stuart, 1991), the *trans*-splicing of nuclear transcripts (Borst, 1986), the existence of the long polycistronic mRNAs (Michels, 1987), the DNA rearrangements associated with antigenic variation in African trypanosomes (Michels, 1987), the postulated transcription of protein-coding genes by RNA poII (Rudenko et al., 1991), the persistence of an intact nuclear envelope during mitosis (Solari, 1980), a highly diverged

Correspondence to: Dr. G. Bernardi, Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France. Tel. (33-1) 4329-5824; Fax (33-1) 4427-7977; e-mail: Bernardi@Arthur.Citi2.FR

*Permanent adresses: (H.M.) Departamento de Bioquímica, Facultad de Ciencias, Tristán Narvaja 1674, Montevideo 11200, Uruguay. Fax (589-2) 409-973; (H.M. and H.R.-M.) Departamento de Genética, Facultad de Medicina, Gral. Flores 2144, Montevideo, Uruguay. Fax (598-2) 949-563.

Abbreviations: aa, amino acid(s); BAMD, bis(aceto-mercurimethyl)dioxane; bp, base pair(s); *ESAG*, expression site-associated gene; GC, % of guanine+cytosine; HK, housekeeping; kb, kilobase(s) or 1000 bp; Myr, 10^6 years; N, any nucleoside; nt, nucleotide (s); *PRO*, gene(s) encoding procyclin(s); R, purine (A or G); *T.*, *Trypanosoma*; *VSG*, gene encoding variable surface glycoprotein (VSG).

SSDI 0378-1119(93)E0727-U

N-terminal H4 histone in *T. cruzi* (Toro et al., 1992), and a physically and enzymatically fragile chromatin in comparison to other eukaryotes (Hecker and Gander, 1985).

On the other hand, comparatively little information is available on the compositional properties of the coding sequences from the nuclear genomes, particularly in *T. cruzi*. It was noted, however, that the base frequencies and codon usage of different classes of genes encoding variable surface glycoproteins, *VSG*s; expression site associated genes, *ESAG*s; and housekeeping genes of *T. brucei*, are different (Michels, 1986; Parsons et al., 1991). This could reflect a genome compartmentalization (see next paragraph below), which in turn might be important for understanding some of the rather unconventional ways used by these parasites in the regulation of gene expression and in development (for reviews, see Michels, 1987; Clayton, 1992).

To gain new insights on this aspect of the genome organization, we analyzed the coding sequences of *T. brucei* and *T. cruzi*. Our results indicate that the compositional patterns and the codon usage and the aa encoding of the two genomes are strikingly different, although in both cases they indicate a compositional compartmentalization which had been previously shown (Isacchi et al., 1993) by the fractionation of the DNA of *T. brucei* (and *T. equiperdum*). This difference between the two genomes is accompanied by a very large divergence in nt at the third codon position among homologous genes. The implications of these results are discussed.

RESULTS AND DISCUSSION

(a) Sequence analysis

The sequences analyzed were obtained from Release 74 of GenBank, and the ACNUC retrieval system (Gouy et al., 1984) was used. The accession Nos. and mnemonics of all genes analyzed are available upon request. Sequences were classified according to the definitions and keywords given by the authors. All non-antigen and non-*ESAG* (see next paragraph) sequences were considered as housekeeping coding sequences.

The 63 sequences analyzed from *T. brucei* comprise 41 housekeeping genes, ten *VSG*s, encoding antigens (which completely cover the trypanosome surface when the parasite is in the bloodstream of the mammalian host), five *PRO* genes (encoding procyclins, i.e., antigens expressed only in the insect) and seven *ESAG*s (genes comprised within the large, 50 kb or so, *VSG* transcription unit; see Michels, 1987). In the case of *T. cruzi*, the analysis was done on 34 coding sequences, 20 housekeeping and 14 antigen sequences.

In the statistical analysis and the construction of codon and aa usage tables, only one *VSG* and one *PRO* were taken into consideration because of the high homology among members of these sequence families.

(b) Compositional distribution of exons and codon positions

Fig. 1 shows the histograms of the compositional distribution of all available coding sequences from *T. brucei* and *T. cruzi* and of their codon positions. *T. brucei* exons cover the 40–60% GC range. In this distribution, *ESAG*s are GC-poor, *VSG*s occupy the central zone, *PRO* s are located at the GC-rich end (see also below), whereas housekeeping genes extend all over the range. These results confirm the trends previously described for *T. brucei* in an analysis covering a more limited number of genes (Michels, 1986). In the case of *T. cruzi*, exons, apart from one housekeeping gene (trbcbpb, coding for a Ca^{2+}-binding protein) with an extremely low GC value (39%), sequences show a distribution covering a GC range comprised between 45% and 67.5%; housekeeping genes appear to be clustered on a slightly higher value compared to the antigen-encoding genes.

The GC content of first and second codon positions from *T. brucei* extends from 40 to 65%, but while housekeeping sequences, *VSG*s and *ESAG*s cover the 40–55% GC range, procyclins display much higher values, 62.5–65%. This difference is due to the fact that *PRO*s have a very biased aa usage: Glu (GAR) being 19.3%, Pro (CCN) 19.2%, Ala (GCN) 14.3%, Gly (GGN) 9.6%. Taken together, these aa represent more than 60% of all the aa. *ESAG*s are rather similar to each other and have a GC content of 40–42.5%. *VSG*s, on the other hand, cover a 5% range, from 47.5% to 52.5%. Housekeeping sequences are clustered between 45% and 50%. In contrast, *T. cruzi* extends from 42.5% to 62.5%, both with housekeeping and antigen sequences.

Third codon positions display the most striking differences between the two species. *T. brucei* extends from 35% to 72.5% in a bimodal distribution with a major peak at 52.5% and a minor peak at 67.5%. The distribution of genes is strikingly non-uniform. While *VSG*s, *ESAG*s and *PRO* genes tend to be located in the GC-poor region, housekeeping sequences extend towards the GC-rich region. This is particularly evident in *PRO*s: the five sequences analyzed exhibit the lowest values. In this connection, similar results were obtained in plant genomes for genes of seed storage proteins (Montero et al., 1990), that also display an extremely biased aa usage, responsible for a very high GC level in first and second codon positions; like *PRO*s, they tend to have low GC values in third codon positions.

T. cruzi, on the other hand, shows a clear bimodal distribution which covers a much more extended region

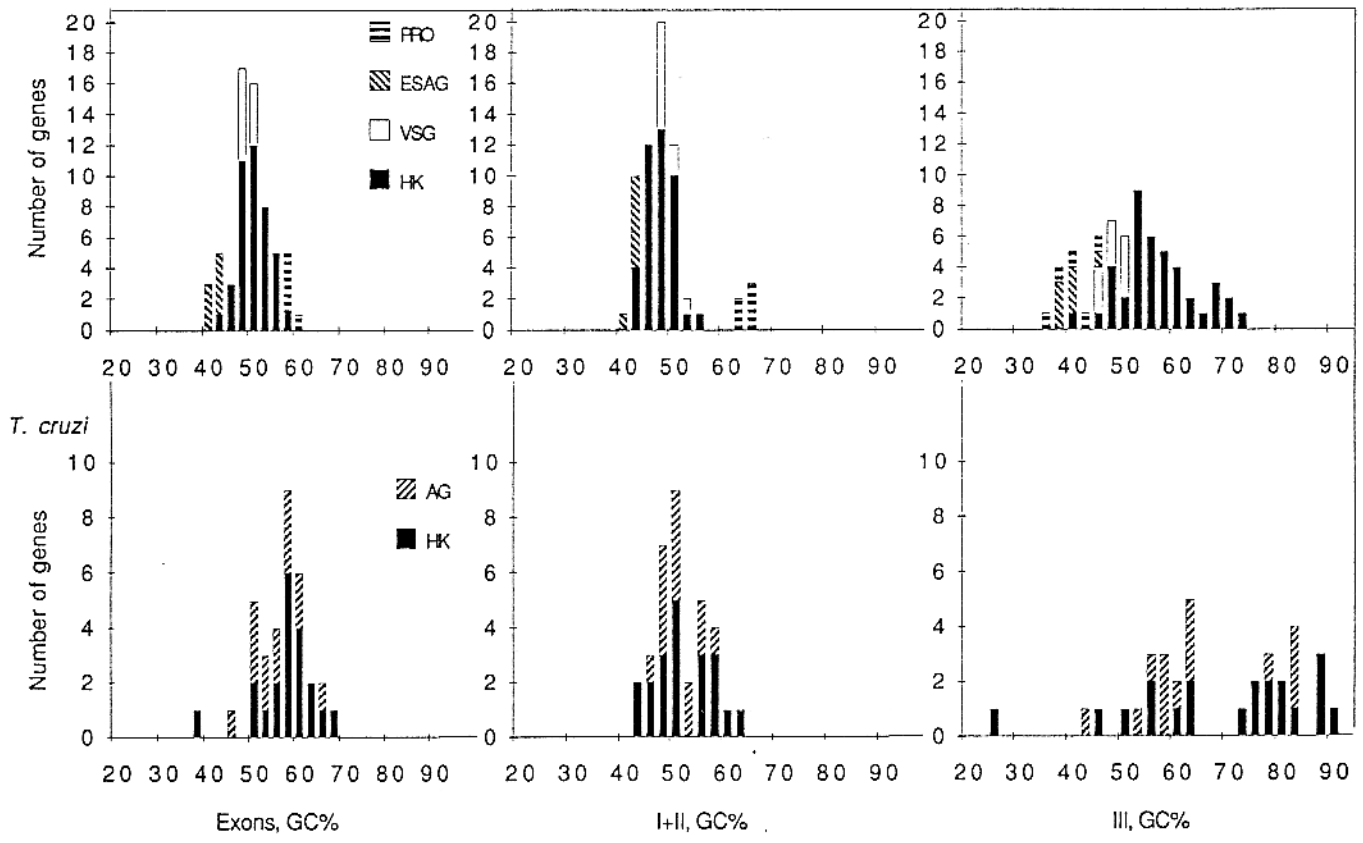

Fig. 1. Compositional patterns of *T. brucei* and *T. cruzi*. Histograms show the distributions of GC levels of exons, first and second, and third codon positions. Ordinates display the number of genes and abscissae the GC levels. *HK*, housekeeping genes; *PRO*, procyclins; *VSG*, variable surface glycoproteins; *ESAG*, expression site associated genes; *AG*, antigens.

(between 42.5% and 90%, if an exceptionally low value, 25%, is neglected). The two groups of values, between 42.5% and 62.5% and between 72.5% and 90%, respectively, are separated by a gap. Both groups contain a similar number of sequences (17 and 16, respectively), but the first one contains more antigen than housekeeping sequences, while the reverse is true in the second. In addition, housekeeping sequences display the extreme values (25% to 90%), while antigen sequences cover the 42.5–82.5% range.

(c) Comparison of GC levels in exons and codon positions

To investigate further the differences described in the previous section, we compared the mean GC values of exons, first, second and third codon positions. Table I shows this analysis both for housekeeping genes and for all the sequences. Antigen sequences alone were not considered because they usually display biased aa usage, and because they are expected to be different in different organisms. In the case of all genes, only one *VSG* and one *PRO* were considered (the longest sequences) because of their high degree of homology. The differences of the mean values were evaluated by the *t*-test, and in the six cases analyzed they were highly significant, the GC content of *T. cruzi* being always higher than that found for *T. brucei*. The comparison of the GC ratio III/(I+II) shows that this value is higher in *T. cruzi* both for housekeeping genes and for all sequences, as expected from the histograms of Fig. 1.

(d) Comparison of homologous sequences

The differences found so far between the two genomes might be due to differences in the gene samples used. That this is not the case is shown by Fig. 2, in which the GC levels of first, second and third codon positions of the eight available pairs of homologous genes are plotted against each other. The GC levels of first and second codon positions are very similar, as expected by the strong conservation of the aa sequences, while in third codon position six out of eight *T. cruzi* genes are richer in GC (by 20 to 30%) than the *T. brucei* homologs, the two similar values being the lowest ones; expectedly, a similar pattern is found when exons are compared (not shown). These results confirm the previous analysis, and suggest that an important shift in the GC levels of third codon positions occurred between the two species since their common ancestor. This shift seems not to have involved genes that are low in third codon position GC,

TABLE I

GC level (%) in exons and in I+II and III codon positions

	Exons (% GC[b])	I+II (% GC)	III (% GC)	III/(I+II)[c]
Housekeeping genes[a]				
T. brucei (*41*)	51.49 (3.35)	48.39 (2.69)	57.72 (7.46)	1.19
T. cruzi (*20*)	58.60 (6.24)	52.55 (5.55)	70.70 (17.20)	1.35
All genes[a, d]				
T. brucei (*50*)	50.29 (4.53)	48.06 (4.08)	54.75 (9.47)	1.14
T. cruzi (*34*)	57.61 (5.73)	52.21 (4.89)	68.43 (15.53)	1.31

[a]The number of genes analyzed in each case are given in parentheses.
[b]Standard deviations are given in parentheses.
[c]I, II, III are GC in codon positions 1–3, respectively.
[d]All genes = housekeeping genes + antigen sequences.

but a larger sample of homologous genes should be tested before reaching a final conclusion on this point.

(e) Comparison of codon usage

Table IIA shows the summary of the differences in codon usage of all genes (housekeeping and antigen) for the two species: 33 (53.2%) codons differ significantly between the two species (ATG (Met) and TGG (Trp) were not compared as there is no degeneracy). To test if this difference is due to antigen genes, the same analysis was done just for housekeeping genes, but the result was very similar since 28 (45.2%) codons differ between the two species.

While this comparison may be criticized because of the different sets of genes used, the large differences found in third codon positions of homologous genes (Fig. 2) supports the conclusion that the two genomes are remarkably different in codon usage.

(f) Analysis of genome compartments

The compositional distribution of sequences displayed in the histograms of Fig. 1 suggests that these genomes are compartmentalized, as indicated by compositional fractionation studies (Isacchi et al., 1993) on *T. brucei* (and *T. equiperdum*). Indeed, in *T. brucei* there is not only a tendency to bimodality in third codon position, but the distribution of coding sequences is different. This trend is also seen, although at a lower degree, in *T. cruzi* (see section **a**) which exhibits a strong bimodality in the GC levels of third codon positions.

To determine whether these observations extend to other levels, the sequences from both genomes were divided according to GC levels in third codon position, and codon usage was determined for each 'compartment' and compared. In *T. cruzi* the cut was done at the 'gap', and in *T. brucei* at a GC level of 52.5%. The latter value was chosen, since 50% is the highest GC level reached by non-housekeeping sequences. The two compartments so defined in each genome are indicated as H (for high GC) and L (for low GC). Again, the analysis was done just for housekeeping genes in order to avoid the biases associated with the aa composition of antigens.

Table IIB shows the codon usage of the two 'compartments' in both *T. brucei* and *T. cruzi*. It can be seen that 26 codons (41.6%) display significant differences in usage in the former, 24 codons (38.7%) in the latter.

(g) Conclusions

Our results can be summarized as follows:

(*1*) The GC levels of third codon positions of the two species differ greatly, being higher in *T. cruzi* than in *T. brucei*; a comparison of the GC levels of eight homologous pair of genes shows the same trend, although the situation may be different for genes having low GC levels in third codon positions; the standard deviations in all codon positions of *T. cruzi* was not only much greater than in *T. brucei*, but even greater than those displayed by avian and mammalian genomes (D'Onofrio and Bernardi, 1992) which are very much larger in size. In turn, this suggests that the genome of *T. cruzi* might be highly compartmentalized, which is in line with its strong bimodality in third codon position, a tendency also shown by *T. brucei* [see also (*4*) below].

(*2*) A strong non-uniformity in the distribution of sequences is evident in *T. brucei*, *VSG*s, *ESAG*s and *PRO*s tending to be located in the GC-poorest region and housekeeping genes extending towards the GC-richest part of the distribution; a similar, although less striking, trend is shown by *T. cruzi*.

(*3*) The comparison of codon usage between the two species shows that roughly half of the codons are differently used. Since both organisms are classified in the same Genus, and their life-cycles are rather similar, the large differences between them are striking; and par-

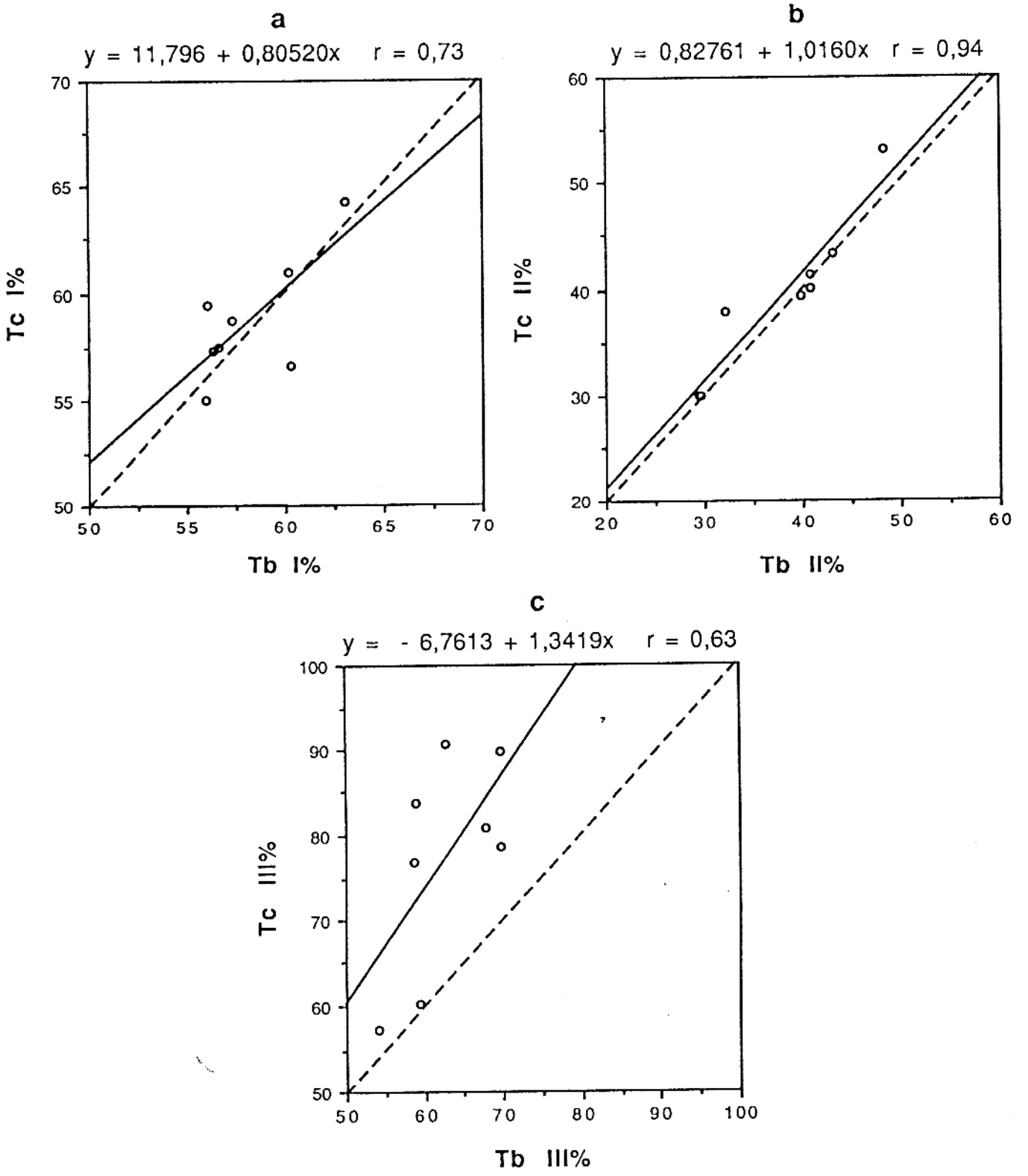

Fig. 2. Comparison of GC level of eight pairs of homologous genes in first (a), second (b), and third (c) codon positions. Ordinates and abscissae display *T. cruzi* (Tc) and *T. brucei* (Tb) values, respectively. The equations for the least-square lines and the correlation coefficients are given. The diagonal line (slope = 1), corresponding to identical values for each pair of genes, is indicated by a dashed line.

ticularly intriguing is the strong shift in GC content. Indeed, codon usage is non-random and was originally considered to be species-specific (Grantham et al., 1981), with similar codon preferences among taxonomically related species (Ikemura, 1985; Anderson and Kurland, 1990). On this basis, one could expect similar biases in

TABLE II

Differences in codon usage

	Species[a]	Genes[b]	Codons[c]	
			No.	%
A	*Tb-Tc*	All genes	33	53.2
	Tb-Tc	Housekeeping genes	28	45.2
B	*Tb*	HHK-LHK	26	41.9
	Tc	HHK-LHK	24	38.7

[a] *Tb*, *T. brucei*; *Tc*, *T. cruzi*.

[b] HHK, high-GC housekeeping genes; LHK, low-GC housekeeping genes.

[c] Number (No.) and percentage (%) of codons for which the mean values in the two genomes or of genome 'compartments' differ significantly.

species belonging to the same genus. However, clearly this is not the case in *T. cruzi* and *T. brucei*, in agreement with the notion that codon usage is highly correlated with genome composition (Bernardi and Bernardi, 1986). As suggested by Gómez et al. (1991), sequence data inferences should be incorporated in the current classification criteria, which is based mainly on morphological stages and host-ranges (McGhee and Cosgrove, 1980), that may or may not reflect evolutionary relationships. At all levels analyzed, our data strongly differentiate both species. It has been postulated that the two species diverged around 100 Myr ago on the basis of comparison of mitochondrial rRNA sequences (Lake et al., 1988). This proposal is currently under further investigation using protein-encoding nuclear sequences. These have already shown enormous levels of silent substitutions (data not shown).

(4) Both genomes seem to be compartmentalized although the compositional heterogeneity is different. As already mentioned, the compartmentalization of the genomes from *T. brucei* (and *T. equiperdum*) was previously detected (Isacchi et al., 1993) by DNA fractionation in preparative Cs_2SO_4/BAMD density gradient [BAMD is bis (acetato mercurimethyl) dioxane]. The interest of this result, showing a striking bimodality, is that it suggests that compartmentalization concerns large regions of relatively homogeneous composition, similar to the isochores of the vertebrate genomes (see Bernardi, 1989 and 1993a,b, for reviews). It should be noted that the case of Trypanosomes is not unique, since compositional compartmentalization has also been demonstrated in the nuclear DNA of *Plasmodium cynomolgi*, which consists of isochores likely to average 100 kb (McCutchan et al., 1988). In conclusion, compositional compartmentalization appears to be a phylogenetically very widespread situation, at least in eukaryotes, as already predicted some years ago (Bernardi et al., 1985).

Note: The Editor of this paper drew our attention to an article (Alonso et al., 1992) comparing base composition and codon usage of smaller sets of coding sequences from *T. brucei* and *T. cruzi*.

ACKNOWLEDGEMENTS

H.M. and H.R.M. thank the UNESCO/Third World Academy of Sciences for short-term fellowships. H.M. thanks too the CSIC Committee of the Universidad de la República, Uruguay, for financial support.

REFERENCES

Alonso, G., Guevara, P. and Ramirez, J.L.: Trypanosomatidae codon usage and GC distribution. Mem. Inst. Oswaldo Cruz, Rio de Janeiro 87 (1992) 517–523.

Anderson, S.G.E. and Kurland, C.G.: Codon preferences in free-living microorganisms. Microbiol. Rev. 54 (1990) 198–210.

Bernardi, G.: The isochore organization of the human genome. Annu. Rev. Genet. 23 (1989) 637–661.

Bernardi, G.: The vertebrate genome: isochores and evolution. Mol. Biol. Evol. 10 (1993a) 186–204.

Bernardi, G.: The isochore organization of the human genome and its evolutionary history – a review. Gene 135 (1993b) 57–66.

Bernardi, G. and Bernardi, G.: Compositional constraints and genome evolution. J. Mol. Evol. 24 (1986) 1–11.

Borst, P.: Discontinuos transcription and antigenic variation in trypanosomes. Annu. Rev. Biochem. 55 (1986) 701–732.

Clayton, C.: Developmental regulation of nuclear gene expression in *Trypanosoma brucei*. Progr. Nucleic Acid Res. Mol. Biol. 43 (1992) 37–66.

de Raadt, P. and Seed, J.R.: Trypanosomes causing disease in man in Africa. In: Kreier, J.P. (Ed.), Parasitic Protozoa, Vol. I. Academic Press, New York, 1977, pp. 175–237.

D'Onofrio, G. and Bernardi, G.: A universal compositional correlation among codon positions. Gene 110 (1992) 81–88.

Fife, E. H.: *Trypanosoma (Schizotrypanum) cruzi*. In: Kreier, J. P. (Ed.), Parasitic Protoza, Vol. I. Academic Press, New York, 1977, pp. 135–173.

Gómez, E., Valdéz, A.M., Piñeiro, D. and Hernández, R.: What is a genus in the Trypanosomatidae family? Phylogenetic analysis of two small rRNA sequences. Mol. Biol. Evol. 8 (1991) 254–259.

Gouy, M., Milleret, F., Mugnier, C., Jacobzone, M. and Gautier, C.: ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res. 12 (1984) 121–127.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R.: Codon catalogue usage is a genomic strategy modulated for gene expressivity. Nucleic Acids Res. 9 (1981) r43–r74.

Hecker, H. and Gander, E.S.: The compaction pattern of the chromatin of *Trypanosomes*. Biol. Cell 53 (1985) 199–208.

Ikemura, T.: Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2 (1985) 13–34.

Isacchi, A., Bernardi, G. and Bernardi, G.: Compositional compartmentalization of the nuclear genomes of *Trypanosoma brucei* and *Trypanosoma equiperdum*. FEBS Lett. 335 (1993) 181–183.

Lake, J.A., de la Cruz, V.F., Ferreira, P.C.G., Morel, C. and Simpson, L.: Evolution of parasitism: kinetoplastid protozoan history reconstructed from mitochondrial rRNA gene sequences. Proc. Natl. Acad. Sci. USA 85 (1988) 4779–4783.

McCutchan, T.F., Dame, J.B., Gwadz, R.W. and Vernick, K.D.: The genome of *Plasmodium cynomolgi* is partitioned into separable domains which appear to differ in sequence stability. Nucleic Acids Res. 16 (1988) 4999–4510.

McGhee, R.B. and Cosgrove, W.B.: Biology and physiology of the lower trypanosomatidae. Microbiol. Rev. 44 (1980) 140–173.

Michels, P.A.M.: Evolutionary aspects of Trypanosomes: analysis of genes. J. Mol. Evol. 24 (1986) 45–52.

Michels, P.A.M.: Genomic organization and gene structure in African trypanosomes. In: Kinghorn, J.R. (Ed.), Gene Structure in Eukaryotic Microbes. IRL Press, Oxford, 1987, pp. 243–262.

Montero, L.M., Salinas, J., Matassi, G. and Bernardi, G.: Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res. 18 (1990) 1859–1867.

Opperdoes, F.R.: Biochemical peculiarities of Trypanosomes, African and South American. Br. Med. Bull. 41 (1985) 130–136.

Parsons, M., Stuart, K. and Smiley, B.L.: *Trypanosoma brucei*: analysis of codon usage and nucleotide composition of nuclear genes. Exp. Parasitol. 73 (1991) 101–105.

Rudenko, G., Chung, H.-M., Pham, V.P. and Van der Ploeg, L.H.T.: RNA pol I can mediate expression of CAT and neo protein-coding genes in *Trypanosoma brucei*. EMBO J. 10 (1991) 3387–3397.

Simpson, L.: Kinetoplast DNA on trypanosomatid flagellates. Int. Rev. Cytol. 99 (1986) 119–179.

Solari, A.J.: The 3-dimensional fine structure of the mitotic spindle in *Trypanosoma cruzi*. Chromosoma 78 (1980) 239–255.

Soulsby, E.J.L.: Helminths, arthropods and protozoa of domesticated animals, 7th Ed. Bailliere Tindall, London, 1982.

Stuart, K.: RNA editing in mitochondrial mRNA of trypanosomatids. Trends Biochem. Sci. 16 (1991) 68–72.

Toro, G.C., Wernstedt, C., Medina, C., Jaramillo, N., Hellman, U. and Galanti, N.: Extremely divergent histone H4 sequence from *Trypanosoma cruzi*: evolutionary implications. J. Cell. Biochem. 49 (1992) 266–271.