FEBS Letters 581 (2007) 5751–5758

Comparative codon and amino acid composition analysis of Tritryps-conspicuous features of *Leishmania major*

Ipsita Chanda[a], Archana Pan[b], Sanjoy Kumar Saha[a], Chitra Dutta[a,*]

[a] *Department of Structural Biology and Bioinformatics, Indian Institute of Chemical Biology, Kolkata 700 032, India*
[b] *Computational Biology Group, Theoretical Physics Department, Indian Association for the Cultivation of Science, Kolkata 700 032, India*

Received 12 July 2007; revised 11 October 2007; accepted 9 November 2007

Available online 26 November 2007

Edited by Takashi Gojobori

Abstract Comparative analyses of codon/amino acid usage in *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi* reveal that gene expressivity and GC-bias play key roles in shaping the gene composition of all three parasites, and protein composition of *L. major* only. In *T. brucei* and *T. cruzi*, the major contributors to the variation in protein composition are hydropathy and/or aromaticity. Principle of Cost Minimization is followed by *T. brucei*, disregarded by *T. cruzi* and opposed by *L. major*. Slowly evolving highly expressed gene-products of *L. major* bear signatures of relatively AT-rich ancestor, while faster evolution under GC-bias has characterized the lowly expressed genes of the species by higher GC_{12}-content.

Keywords: Gene expressivity; GC-bias; Hydropathy; Aromaticity; Correspondence analysis; Principle of cost minimization

1. Introduction

The trypanosomatid pathogens *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*, often referred together as "Tritryps" [1], are three closely related kinetoplastid parasitic protozoa that cause some of the most debilitating diseases of humankind – cutaneous leishmaniasis, African sleeping sickness and Chagas disease, respectively [2]. All three parasites possess complex life-cycles alternating between the specific insect vectors and the mammalian hosts, undergoing distinct developmental changes in the insect vectors [3–5] that allow them to infect the human host. In spite of considerable research efforts, no vaccine could be approved yet for any of the diseases caused by these pathogens and the drugs in use are highly toxic [4] and prone to the development of drug resistance [6]. There is, therefore, an urgent need to understand the biology of these pathogens and people are trying to exploit their genome information [3–5] in this regard. *L. major*, *T. brucei* and *T. cruzi* contain about 32.8, 26 and 55-megabase size haploid genomes distributed in 36, 11 and 28 chromosomes with an average GC-content of 59.7%, 46.4% and 51%, respectively. Comparative analyses [1] revealed that the three genomes share 6158 ortholog clusters of protein-coding genes, which exist in large syntenic blocks containing 80% of the *T. brucei* and 93% of the *L. major* genes. They also share a number of molecular and biochemical characters [7]. Yet the Tritryps differ in features like mode of transmission by different insects, different target tissues, distinct disease pathogenesis and use of different strategies of immune evasion [1]. In *L. major* genes, a negative correlation exists between GC_{12} and GC_3, the origin of which has remained an open question [8]. For *T. brucei* and *T. cruzi*, however, this correlation is positive. In an effort to analyze the compositional similarities and divergence within and across these genomes in further details, we report a comparative multivariate analysis of their codon and amino acid usage patterns.

*Corresponding author. Fax: +91 33 2473 0284/5197.
E-mail addresses: ipsita_chanda@yahoo.co.in (I. Chanda), archanpan@gmail.com (A. Pan), sanjoy574@yahoo.co.in (S.K. Saha), chitradutta@hotmail.com, cdutta@iicb.res.in (C. Dutta).

Abbreviations: ORF, Open reading frame; RSCU, relative synonymous codon usage; CAI, codon adaptation index; RAAU, relative amino acid usage; GC_{3S}, G + C content at synonymous codon sites excluding ATG for Met and TGG for Trp; GC_{12}, G + C content at first and second codon sites; COA, correspondence analysis; VSG, variable surface glycoprotein; DGF-1, dispersed gene family protein -1; MMW, mean molecular weight; PCM, principle of cost minimization

2. Materials and methods

2.1. Genome sequence data

The nuclear genome sequence of *L. major* with 8272 protein-coding genes was extracted from Sanger database (http://www.sanger.ac.uk/) and those of *T. cruzi* and *T. brucei* with 12570 and 9068 from TIGR Database (http://www.tigr.org). Annotations of the open reading frames (ORFs) were cross-checked with GeneDB. To reduce the sampling error, the genes with less than 100 codons, internal stop codons, untranslated codons and pseudogenes were excluded from the analysis, resulting in the datasets of 7806, 6084 and 11627 predicted ORFs for *L. major*, *T. brucei* and *T. cruzi*, respectively.

2.2. Parameters used to identify the trends of variations within protein-coding genes

For each ORF/ORF-products under study, the following parameters were calculated: relative synonymous codon usage (RSCU), codon adaptation index (CAI) [9], the G + C content at synonymous codon sites excluding ATG for Met and TGG for Trp (GC_{3S}), relative amino acid usage (RAAU), G + C content at first and second codon sites (GC_{12}), average hydropathy [10], Aromaticity [11] and Alcoholicity [12] of the gene-products.

2.3. Datasets of highly and lowly expressed genes

Datasets of putative highly and lowly expressed genes were prepared taking genes from the two extreme ends of Axis1 of correspondence analysis (COA) on RSCU in all the three parasites (Supplementary Table 1). Highly expressed genes were characterized by high codon

0014-5793/$32.00 © 2007 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2007.11.041

adaptation index (CAI) (most of them being experimentally characterized house-keeping genes), whereas the lowly expressed genes were characterized by low CAI values (Supplementary Fig. 1).

2.4. Statistical analyses

Most analyses were performed using the program CodonW 1.4.2 (http://molbiol.ox.ac.uk/win95.codonW.zip). COA [13] was used to explore the variation of RSCU values and amino acid usage. In order to detect the significant differences in codon and amino acid usage, 2×2 contingency table χ^2 method was used.

2.5. Estimation of non-synonymous and synonymous substitutions in highly and lowly expressed genes

About 50 1:1:1 orthologs each for different species of Leishmania (e.g., *L. donovoni*, *L. braziensis*, *L. infuntum*, etc.), *T. brucei* and *T. cruzi* were extracted using BLASTP for the potential highly expressed and lowly expressed genes (lying at the two extremity of Axis1 of COA on RSCU) of *L. major*. The homologs with e-values <e−50 were considered as orthologs. Pairwise alignments between the orthologs and the estimation of the number of synonymous substitutions per synonymous site, d_S and non-synonymous substitutions per non-synonymous site, d_N were carried out using ClustalW (with default settings) and MEGA program (version 2.1) [14], respectively. Comparisons of the substitution pattern between the datasets of highly and lowly expressed genes were done using Kolmogorov–Smirnov statistical test.

3. Results

3.1. Major sources of variations in synonymous codon usages in the three parasites

To identify the major sources of intra-species variations in synonymous codon preferences in the three parasites, COA on RSCU has been performed on *L. major*, *T. brucei* and *T. cruzi* datasets, respectively. As shown in Table 1, Axis1 accounts for 16.59%, 10.71% and 13.23% of the total variations for RSCU in *L. major*, *T. brucei* and *T. cruzi*, respectively. In all cases, Axis1 exhibits strong correlations with CAI and GC_{3S}, suggesting that the translational selection [9], along with directional mutational pressure [15], play a major role in governing the synonymous codon usage. In *L. major*, Axis2 exhibits significant correlation with GT_{3S} of the genes only, but in *T. brucei* and *T. cruzi*, Axis2 is correlated not only with GT_{3S} of the genes, but also with the mean hydropathy, aromaticity and Thr-content of the gene-products (Table 1).

In the Axis1–Axis2 plot of each genome under study, the highly expressed genes are clustered at one end of Axis1 (Fig. 1a–c, red), indicating that these genes follow a distinct pattern of synonymous codon usage. A comparison of RSCU values of the highly expressed genes with those of the lowly expressed genes shows that in all three parasites under study, a similar subset of synonymous codons, mostly G-/C-ending, (Table 2, bold letters) are preferred by the highly expressed genes. In *T. brucei* and *T. cruzi*, the lowly expressed genes exhibit relatively higher usage of A-/U-ending codons. But in *L. major*, even the lowly expressed genes prefer to use G-/C-ending codons for most of the amino acids, though the frequencies of such codons are lower than those in the highly expressed genes. This is in agreement with the higher GC-content of the *L. major* genome (59.7%). As seen in Table 2, the extent of bias in the synonymous codon usage is highest in *L. major* and lowest in *T. brucei,* suggesting that among the three species, the influence of translational selection is strongest in *L. major*.

3.2. Distinct codon usage in variant surface glycoproteins (VSG) in T. brucei and dispersed gene family protein-1 (DGF-1) in T. cruzi

All genes other than the highly expressed ones in *L. major* constitute a single cluster in Fig. 1a, indicating that they follow similar codon usage patterns. But in *T. brucei* (Fig. 1b), there are three distinct clusters formed by (a) the highly expressed genes (red), (b) the variant surface glycoproteins or VSG genes (blue) and (c) the rest of genes (black). In *T. brucei*, the key to survival is a huge repertoire of antigenically distinct VSGs, expression patterns of which change periodically during a chronic infection [16]. Switching the expressed VSG allows the parasite population to escape immune killing mediated by the antibodies produced against the previously expressed VSG [17]. Segregation of VSGs (blue) in Fig. 1b indicates that the synonymous base usage in VSGs is distinct from that in other genes. Positive correlation of GC_{3S} and CAI with Axis1 (Table 1) suggests that VSGs are characterized by relatively low GC_{3S} and CAI values (Fig. 1b), while the positive correlation of GT_{3S} with Axis2 (Table 1) implies significantly low usage of G_3/T_3 in VSG genes (Supplementary Table 2).

T. cruzi does not use the strategy of antigenic variation for host immune evasion, it rather exhibits a variable repertoire of surface molecules and the highly polymorphic antigenic components that represent a useful arsenal for host cell invasion. The surface of *T. cruzi* is covered by different groups of carbohydrate-rich mucin-like glycoproteins/mucin Tc MUCII

Table 1
Major trends in synonymous codon usage in *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi* as revealed by COA on RSCU of genes

	Axis1			Axis2		
	Total variability	Source of variation	Correlation coefficient[a] (*r*-value)	Total variability	Source of variation	Correlation coefficient[a] (*r*-value)
L. major	16.59	CAI	−0.96	4.58	GT_{3S}	0.55
		GC_{3S}	−0.95			
T. brucei	10.71	CAI	0.90	5.92	GT_{3S}	0.59
		GC_{3S}	0.85		Gravy	0.28
					Aromaticity	0.27
					Thr-content	−0.27
T. cruzi	13.23	CAI	−0.87	6.68	GT_{3S}	0.83
		GC_{3S}	−0.94		Gravy	0.18
					Aromaticity	0.23
					Thr-content	−0.22

[a]All correlations are significant at $P < 0.0001$.

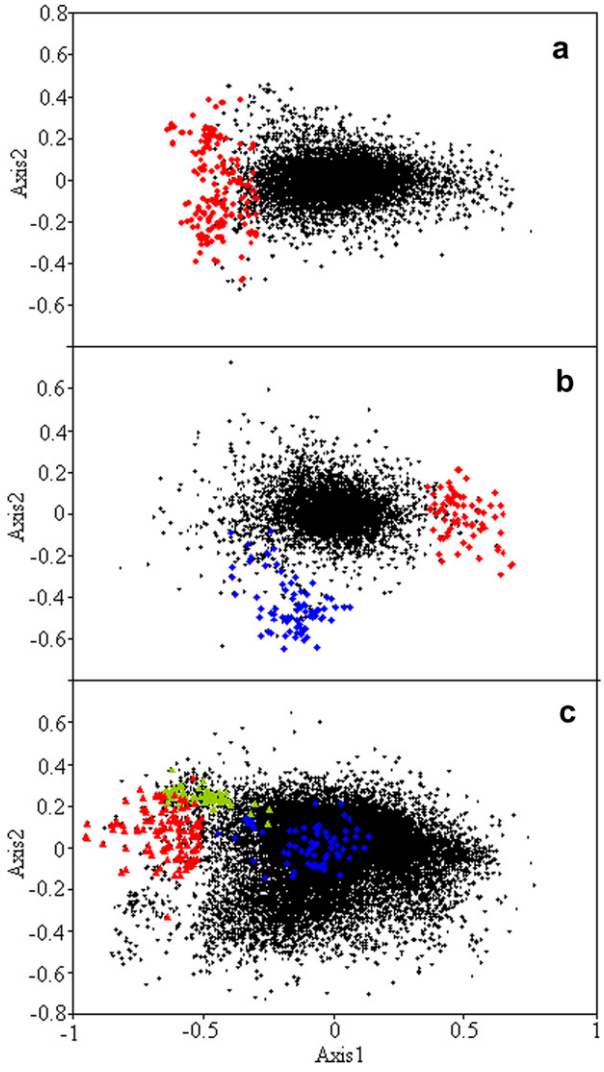

Fig. 1. Position of genes along Axis1 generated by COA on RSCU has been plotted against Axis2 in: (a) *Leishmania major*; (b) *Trypanosoma brucei*; and (c) *Trypanosoma cruzi*. Highly expressed genes, VSG, mucin Tc MUCII proteins and DGF-1 are represented by red, blue and green, respectively.

that are differentially expressed during the mammal-dwelling stages of parasite life cycle [18]. In the Axis1–Axis2 plot of COA on RSCU of *T. cruzi* genes, the mucin Tc MUCII genes (blue) merge with the moderately or lowly expressed genes, indicating that the synonymous codon usage in these genes follows the general trend of the genome (Fig. 1c). However, there is a group of genes, putatively encoding dispersed gene family protein-1 (DGF-1) (green), which appears just above the cluster of highly expressed genes (red) (Fig. 1c). As indicated by the co-segregation of DGF-1 with highly expressed genes in Fig. 1c, DGF-1s are also characterized by high GC_{3S} and high CAI values and their synonymous codon bias is similar to that of the highly expressed genes (Supplementary Table 2), showing thereby a potential for high expression.

3.3. Major sources of variations in amino acid usages – distinct features of *L. major* proteins

In order to identify the major trends of intra-proteomic variations in amino acid composition in Tritryps, COA on amino acid usage has been carried out for each species. The first two axes generated by COA account for 44.48%, 16.63% and 19.91% of the total variations in *L. major*, *T. brucei* and *T. cruzi*, respectively (Table 3). A distinct feature of *L. major* is that GC_{12} and CAI, along with Alcoholicity and Aromaticity constitute the primary sources of intra-proteomic variations in amino acid usage, mean Hydrophathy being the secondary factor. But in *T. brucei* and *T. cruzi*, the proteome composition seems to be dictated, not by gene expressivity or GC-content, but by the physicochemical factors like Hydropathy, Aromaticity or Alcoholicity. In *T. brucei*, Gravy score and Aromaticity, both act as the primary sources of variation, but in *T. cruzi*, Gravy score alone is the primary source of such variation, Aromaticity and Alcoholicity of proteins being the secondary sources (Table 3).

In consistence with these observations, highly expressed genes (red) of *L. major* cluster at the extreme right end of Axis1 in the Axis1–Axis2 plot of COA on amino acid usage (Fig. 2a), whereas in *T. brucei* and *T. cruzi,* highly expressed genes merge with the main cluster of gene-products (Fig. 2b, c). VSGs (blue) of *T. brucei* appear towards the right of the highly expressed genes (Fig. 2b). As can be seen from the Supplementary Table 3, VSGs are characterized by exceptionally high frequencies of Ala, Thr, Asn and Lys and low frequencies of Val, Arg, Met, etc. In *T. cruzi* (Fig. 2c), the cluster of highly expressed genes (red) is well segregated from the mucin Tc MUCII proteins (blue) and the DGF-1 (green), indicating that the highly expressed genes differ appreciably from other two groups of proteins in amino acid composition (Supplementary Table 3).

Fig. 2a and Table 3 together suggest that the highly expressed genes of *L. major* are characterized by relatively low GC_{12}, low Alcoholicity and high Aromaticity. These were not expected because (i) *L. major* is a relatively GC-rich organism with average GC-content 59.7% and (b) according to the principle of cost minimization (PCM) [19], the highly expressed genes of most of the unicellular organisms including parasitic ones [20] often prefer to use residues having low aromaticity and low mean molecular weight (MMW). Our analysis reveals that the CAI values of *L. major* genes exhibit significant positive correlations with Aromaticity and MMW of the respective gene-products ($r = 0.10$ and 0.18, $P < 0.001$, respectively), implying that *L. major* genes not only disregard the PCM, they rather oppose the principle in a sense that the highly expressed genes of this organism selectively use the residues of high bioenergetic cost. In *T. brucei*, CAI values of the genes exhibit negative correlations with MMW and Aromaticity ($r = -0.25$ and -0.15, $P < 0.001$, respectively), implying that the PCM is obeyed by this parasite, but in *T. cruzi*, neither MMW nor Aromaticity bears any significant correlation with CAI.

Figs. 1a and 2a indicate that the highly expressed genes of *L. major* are characterized by lower GC_{12} and higher GC_{3S} as compared to other genes of the species, which is in accordance with the negative correlation between GC_{12} and GC_3 of its genes [8]. Table 4 reveals that GC_1 and GC_2 of the highly expressed genes of *L. major* are similar to those of the highly and lowly expressed genes of *T. cruzi* and *T. brucei*. But GC_1/GC_2 of the lowly expressed genes of *L. major* is significantly higher from GC_1/GC_2 of all other groups of genes of Tritryps (Table 4). Fig. 3 shows the average amino acid frequencies in the highly and lowly expressed genes of three parasites under

Table 2
RSCU values of different groups of genes of *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*

Amino acid	Codon	*L. major*		*T. cruzi*		*T. brucei*	
		HEG[a]	LEG[a]	HEG	LEG	HEG	LEG
Phe	UUU	0.25	1.13[c]	0.93	1.68[c]	0.67	1.42[c]
	UUC	**1.75**[b]	0.87	**1.07**[b]	0.32	**1.33**[b]	0.58
Leu	UUA	0.01	0.31[c]	0.08	1.17[c]	0.20	1.25[c]
	UUG	0.20	1.00[c]	0.71	1.52[c]	0.95	1.36[c]
	CUU	0.38	1.20[c]	0.78	1.68[c]	1.41	1.33
	CUC	1.49[b]	1.28	0.97[b]	0.46	1.49[b]	0.69
	CUA	0.08	0.46[c]	0.08	0.45[c]	0.33	0.54[c]
	CUG	**3.85**[b]	1.76	**3.38**[b]	0.73	**1.63**[b]	0.83
Ile	AUU	0.43	1.27[c]	1.12	1.75[c]	1.26	1.42[c]
	AUC	**2.53**[b]	1.20	**1.60**[b]	0.42	**1.50**[b]	0.46
	AUA	0.05	0.53[c]	0.28	0.83[c]	0.23	1.12[c]
Val	GUU	0.31	0.84[c]	0.60	1.54[c]	0.96	1.38[c]
	GUC	0.90	0.93	0.60[b]	0.49	0.76[b]	0.47
	GUA	0.08	0.54[c]	0.10	0.74[c]	0.41	0.83[c]
	GUG	**2.71**[b]	1.68	**2.70**[b]	1.23	**1.88**[b]	1.33
Ser	UCU	0.54	1.05[c]	0.37	1.59[c]	0.98	1.19[c]
	UCC	1.69[b]	0.85	1.02[b]	0.87	1.34[b]	0.73
	UCA	0.07	0.81[c]	0.33	1.34[c]	0.73	1.14[c]
	UCG	**1.79**[b]	1.18	1.79[b]	0.57	0.99[b]	0.58
	AGU	0.14	0.69[c]	0.51	1.03[c]	0.67	1.45[c]
	AGC	**1.76**[b]	1.42	**1.99**[b]	0.59	**1.29**[b]	0.91
Pro	CCU	0.27	0.95[c]	0.42	1.25[c]	0.78	1.28[c]
	CCC	0.99[b]	0.76	1.02[b]	0.63	**1.32**[b]	0.77
	CCA	0.15	0.97[c]	0.51	1.46[c]	0.89	1.31[c]
	CCG	**2.59**[b]	1.31	**2.04**[b]	0.66	1.01[b]	0.64
Thr	ACU	0.21	0.75[c]	0.36	0.99[c]	0.72	1.11[c]
	ACC	1.19[b]	0.90	0.80[b]	0.71	0.95[b]	0.62
	ACA	0.22	1.18[c]	0.39	1.53[c]	1.00	1.48[c]
	ACG	**2.38**[b]	1.18	**2.45**[b]	0.77	**1.33**[b]	0.78
Ala	GCU	0.45	0.88[c]	0.68	1.09[c]	1.02	1.24[c]
	GCC	1.49[b]	0.95	0.91[b]	0.72	1.18[b]	0.62
	GCA	0.14	1.00[c]	0.40	1.38[c]	0.79	1.35[c]
	GCG	**1.92**[b]	1.17	**2.01**[b]	0.80	**1.01**[b]	0.79
Tyr	UAU	0.09	0.74[c]	0.28	1.31[c]	0.58	1.27[c]
	UAC	**1.91**[b]	1.26	**1.72**[b]	0.69	**1.42**[b]	0.73
His	CAU	0.16	0.69[c]	0.34	1.29[c]	0.53	1.16[c]
	CAC	**1.84**[b]	1.31	**1.66**[b]	0.71	**1.47**[b]	0.84
Gln	CAA	0.05	0.66[c]	0.21	1.23[c]	0.55	1.27[c]
	CAG	**1.95**[b]	1.34	**1.79**[b]	0.77	**1.45**[b]	0.73
Asn	AAU	0.11	0.76[c]	0.36	1.37[c]	0.58	1.22[c]
	AAC	**1.89**[b]	1.24	**1.64**[b]	0.63	**1.42**[b]	0.78
Lys	AAA	0.04	0.67[c]	0.37	1.27[c]	0.46	1.20[c]
	AAG	**1.96**[b]	1.33	**1.63**[b]	0.73	**1.54**[b]	0.80
Asp	GAU	0.36	0.90[c]	0.43	1.41[c]	0.82	1.35[c]
	GAC	**1.64**[b]	1.10	**1.57**[b]	0.59	**1.18**[b]	0.65
Glu	GAA	0.10	0.66[c]	0.35	1.28[c]	0.62	1.19[c]
	GAG	**1.90**[b]	1.34	**1.65**[b]	0.72	**1.38**[b]	0.81
Cys	UGU	0.09	0.69[c]	0.32	1.28[c]	0.62	1.20[c]
	UGC	**1.91**[b]	1.31	**1.68**[b]	0.72	**1.38**[b]	0.80
Arg	CGU	0.60	1.02[c]	0.94	1.48[c]	1.69[b]	0.94
	CGC	**4.82**[b]	1.77	**2.44**[b]	0.64	**2.85**[b]	0.44
	CGA	0.06	0.98[c]	0.44	1.12[c]	0.38	0.79[c]

Table 2 (*continued*)

Amino acid	Codon	*L. major*		*T. cruzi*		*T. brucei*	
		HEG[a]	LEG[a]	HEG	LEG	HEG	LEG
Arg	CGG	0.36	1.03[c]	1.07[b]	0.69	0.66[b]	0.55
	AGA	0.03	0.53[c]	0.21	1.23[c]	0.10	1.81[c]
	AGG	0.14	0.68[c]	0.90[b]	0.83	0.32	1.47[c]
Gly	GGU	0.73	0.99	0.75	1.32[c]	1.64[b]	1.15
	GGC	**2.92**[b]	1.46	**2.07**[b]	0.75	**1.23**[b]	0.57
	GGA	0.08	0.74[c]	0.32	1.24[c]	0.59	1.57[c]
	GGG	0.28	0.81[c]	0.86[b]	0.69	0.54	0.71[c]

Bold letters: The codon optimally used by a particular amino acid residue.
[a]HEG and LEG: Groups of potential highly and lowly expressed genes taken from two extreme ends of axis1 of COA of RSCU of genes in the respective species.
[b]Codons having significantly higher frequencies in HEG compared to LEG ($P < 0.001$).
[c]Codons having significantly higher frequencies in LEG compared to HEG ($P < 0.001$).

Table 3
Major trends in amino acid usage in *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi* as revealed by COA on amino acid usage of the encoded proteins

	Axis1			Axis2		
	Total variability	Sources of variation	Correlation coefficient[a] (*r*-value)	Total variability	Sources of variation	Correlation coefficient[a] (*r*-value)
L. major	30.21	GC_{12}	−0.86	14.27	Aromaticity	0.67
		CAI	0.44		Gravy	0.51
		Alcoholicity	−0.78			
		Aromaticity	0.53			
T. brucei	10.71	Gravy	−0.82	5.92	Alcoholicity	0.56
		Aromaticity	−0.76			
T. cruzi	13.23	Gravy	−0.68	6.68	Aromaticity	−0.78
					Alcoholicity	0.63

[a]All correlations are significant at $P < 0.0001$.

study. Frequencies of many amino acids differ widely in the highly and lowly expressed genes of *L. major* and in some cases, the values come at the two extreme ends (Fig. 3, pink square and triangle). Fig. 3 also shows that the frequencies of residues encoded by AU-rich codons such as Phe, Ile, Tyr, Asn and Lys are significantly lower, but those of Pro, Ala and Ser are higher in the lowly expressed genes of *L. major* than the lowly expressed genes of two trypanosomes.

3.4. Greater conservation of highly expressed genes

Estimation of d_N, d_S and d_N/d_S on the orthologs of highly and lowly expressed genes of *L. major* and other species of Leishmania and those of *L. major–T. brucei*, *L. major–T. cruzi* and *T. brucei–T. cruzi* (Table 5) shows that in all three species, both d_N and d_N/d_S values are significantly lower for the highly expressed genes than the lowly expressed genes, suggesting that the non-synonymous codon positions of the highly expressed genes are more conserved than their lowly expressed counterparts. This means that the amino acid composition of the highly expressed genes of Tritryps is closer to the ancestor. Therefore, a plausible reason of the AT-richness of the highly expressed genes of *L. major* as compared to the lowly expressed genes of the same species could be that they have been derived from a relatively AT-rich ancestor and the lowly expressed genes, being evolved at a faster rate under increasing GC-bias, have become GC-richer than their highly expressed counterparts. The higher GC_1/GC_2-content of the lowly expressed genes of *L. major* as compared to those of *T. brucei* and *T. cruzi,* could be due to stronger mutational bias in *L. major* towards increasing GC. That the GC-bias in *Leishmania* is stronger than that in *Trypanosomes* is apparent from appreciably higher GC_3-content of both highly and lowly expressed genes of the former than those of the later (Table 4).

There is no significant difference in d_S values of the highly and lowly expressed genes in Tritryp lineage. This was unexpected because the highly expressed genes usually exhibit significantly lower d_S than the respective lowly expressed genes in the organisms under translational selection [21].

It is interesting to note that d_S values of *T. cruzi* vs *L. major* is significantly lower than those of *T. brucei* vs *T. cruzi* or *T. brucei* vs *L. major* (Table 6). This means that since their separation from the Leishmania lineage [22], *T. brucei* has deviated at the synonymous codon positions at much faster rate than *T. cruzi*. However, the d_N value of *T. brucei* vs *T. cruzi* is significantly lower than that of *T. brucei* vs *L. major* and *T. cruzi* vs *L. major* (Table 6), indicating that the protein sequences in African and American trypanosomes have not been diverted much since their separation from the common ancestor.

4. Discussion

The present study reveals the major differences between the selection forces shaping the gene/protein composition of Tritryps. In *L. major*, not only the synonymous codon usage, but also the amino acid variation is dictated by mutational bias and translational selection. On contrary, in *T. brucei* and *T. cruzi*, the physicochemical factors like hydropathy or aromaticity govern the amino acid variation solely and even the synonymous codon usage partially (the major contribution to

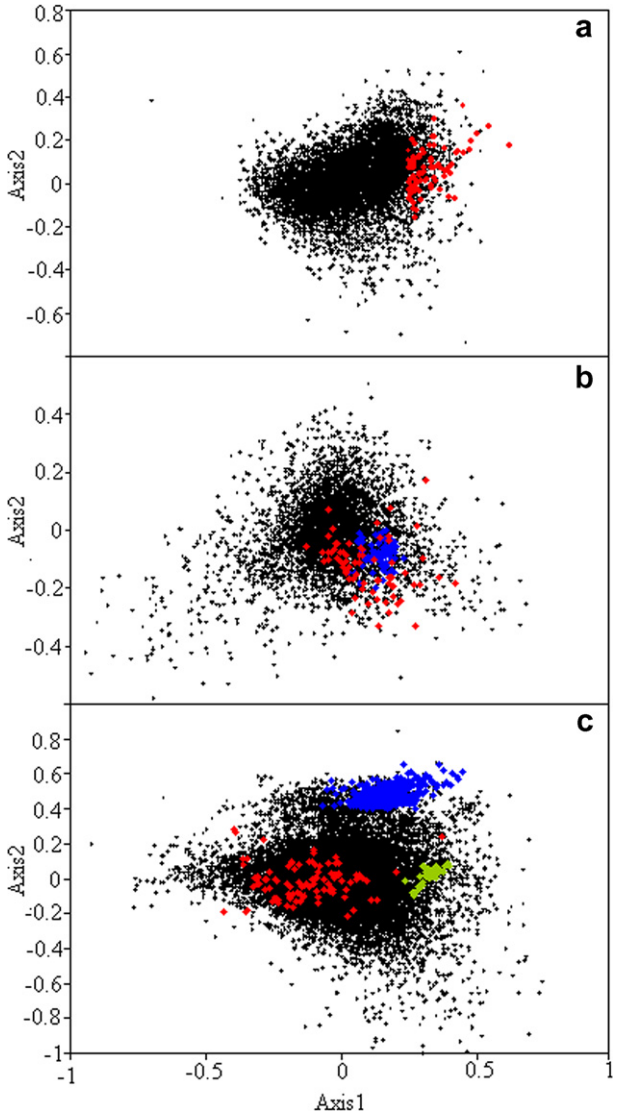

Fig. 2. Position of genes along Axis1 generated by COA on amino acid usage has been plotted against Axis2 in: (a) *Leishmania major*; (b) *Trypanosoma brucei*; and (c) *Trypanosoma cruzi*. Highly expressed genes, VSG, mucin Tc MUCII proteins and DGF-1 are represented by red, blue and green, respectively.

synonymous codon usage, however, come from GC-bias and translational selection).

Lower values of GC_{12}, d_N and d_N/d_S of the highly expressed genes of *L. major*, as compared to their lowly expressed counterparts suggest that the highly expressed gene-products are closer to their ancestral composition, which might have been relatively rich in AT-content, while the lowly expressed gene-products have evolved at faster rate under increasing GC-bias.

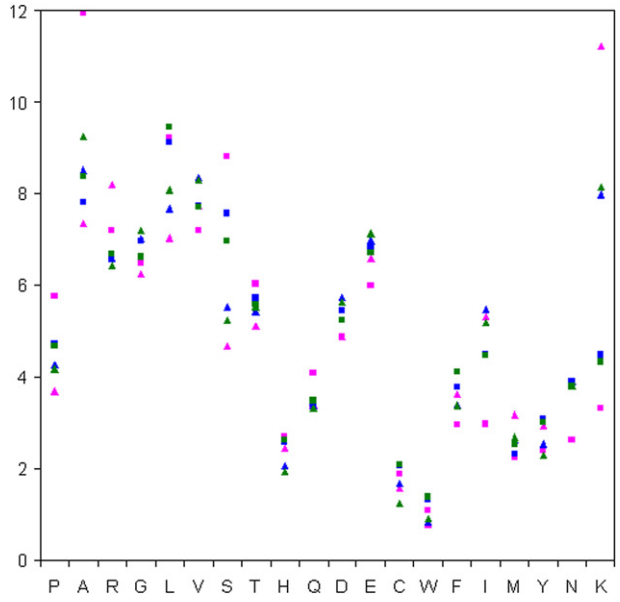

Fig. 3. Amino acid composition of highly and lowly expressed genes of Tritryps. Pink, blue, green triangles and squares represent highly and lowly expressed genes of *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*, respectively.

Due to purifying selection, the GC-bias could not affect much the non-synonymous sites of the highly expressed genes of *L. major*, but as the translational selection acts more strongly on the synonymous sites of the highly expressed genes than that of the lowly expressed genes and as the optimal codons of *L. major* are mostly G-/C-ending, the synonymous sites of the highly expressed genes have evolved towards higher GC-values. As a consequence, the highly expressed genes of *L. major* are characterized by lower GC_{12} and higher GC_3 than their lowly expressed counterparts and probably due to this, a significant negative correlation has been developed between GC_{12} and GC_3 of *L. major* genes [8]. In *T. brucei* and *T. cruzi*, the GC-bias was not strong enough to create a significant difference in GC_{12} composition of the highly and lowly expressed genes. Furthermore, proteins in *L. major* could afford to evolve against the principle of cost minimization, and *T. cruzi* proteins could ignore it, but *T. brucei* has evolved in accordance with the principle. It is, however, not clear why the synonymous sites of the highly expressed genes, which are under translational control, are evolving at almost same rate as the lowly expressed genes in all three organisms under study.

Appreciable differences in codon/amino acid usage patterns also exist among specific groups of genes/gene-products of the African and American trypanosomes. Most interesting among them are the diverse trends in codon and/or amino acid usage in the immunogenic arsenals of the two trypanosomes, i.e., the VSGs of *T. brucei* and mucin Tc MUCII proteins of

Table 4
GC-content of highly and lowly expressed genes of Tritryps at three codon positions

	Highly expressed genes			Lowly expressed genes		
	Leishmania major	*Trypanosoma brucei*	*Trypanosoma cruzi*	*L. major*	*T. brucei*	*T. cruzi*
GC	0.62	0.54	0.57	0.61	0.51	0.51
GC_1	0.58	0.57	0.58	0.61	0.57	0.57
GC_2	0.41	0.40	0.40	0.45	0.42	0.42
GC_3	0.85	0.63	0.73	0.77	0.52	0.55

Table 5
Estimation of d_N, d_S, d_N/d_S between orthologs of highly and lowly expressed genes of *Leishmania major*

	d_N			d_S			d_N/d_S		
	HEG[a]	LEG[a]	D[b]	HEG	LEG	D	HEG	LEG	D
L. major–Leishmania sp.	0.028	0.049	0.361**	0.146	0.204	0.321*	0.203	0.232	0.306*
L. major–Trypanosoma brucei	0.167	0.362	0.596**	0.575	0.558	0.249	0.337	0.771	0.545**
L. major–Trypanosoma cruzi	0.175	0.354	0.666**	0.460	0.501	0.242	0.411	0.830	0.500**
T. brucei–T. cruzi	0.108	0.270	0.667**	0.614	0.630	0.326	0.206	0.537	0.589**

[a]Highly and lowly expressed genes, respectively.
[b]Maximum difference between the cumulative distributions.
*$P < 0.01$ in Kolmogorov–Smirnov test.
**Significance value $P < 0.001$.

Table 6
Estimation of number of synonymous substitutions per synonymous site (d_S) and number of non-synonymous substitutions per non-synonymous site (d_N)

Ortholog pairs	Mean d_S	Mean d_N	Mean d_N/d_S
T. brucei vs *T. cruzi*	0.63	0.21	0.51
T. brucei vs *L. major*	0.60	0.28	0.56
T. cruzi vs *L. major*	0.55	0.27	0.61

T. cruzi. Among the other fascinating observations made in the present study are the significant contributions of DGF-1 to intra-genomic variations in codon/amino acid usages in *T. cruzi*. Frequent occurrence of putative transmembrane domains, ordered globular structure and EGF-like domain signature of DGF-1 (data not shown) suggest that they might have been associated with some essential membrane function, important for creating host–parasite interactions.

Another observation that deserves mention is the higher rate of synonymous substitution between the *T. cruzi–T. brucei* orthologs than that between *T. cruzi–L. major* orthologs. The observation, though unanticipated, is in accordance with the phylogenetic study of 18S rRNA sequences [23], which proposed that since their divergence from the Leishmania lineage, *T. brucei* and the other mammalian tsetse-transmitted trypanosomes might have been evolving several times faster than *T. cruzi* and its relatives.

Acknowledgements: This work was supported by the Council of Scientific and Industrial Research (Project No. CMM 0017) and Department of Biotechnology, Government of India (Grant Number BT/BI/04/055-2001). We are grateful to Prof. J. Chakrabarti, IACS, Kolkata and Dr. Shuvagata Ghosh, IICB, for critical reading of the manuscript and Mr S. Bag for giving technical support for the extraction of *L. major* genome sequence.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2007.11.041.

References

[1] El-Sayed, N.M., Myler, P.J., Blandin, G., et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science 309, 404–409.
[2] Parsons, M., Worthey, E.A., Ward, P.N. and Mottram, J.C. (2005) Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. BMC Genomics 6, 127–145.
[3] El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., et al. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of chagas disease. Science 309, 409–415.
[4] Berriman, M., Ghedin, E., Hertz-Fowler, C., et al. (2005) The genome of the African trypanosome *Trypanosoma brucei*. Science 309, 416–422.
[5] Ivens, A.C., Peacock, C.S., Worthey, E.A., et al. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. Science 309, 436–442.
[6] Borst, P. and Ouellette, M. (1995) New mechanisms of drug resistance in parasitic protozoa. Annu. Rev. Microbiol. 49, 427–460.
[7] Musto, H., Rodríguez-Maseda, H. and Bernardi, G. (1994) The nuclear genomes of African and American trypanosomes are strikingly different. Gene 141, 63–69.
[8] NecÅŸulea, A. and Lobry, J.R. (2006) Revisiting the directional mutation pressure theory: the analysis of a particular genomic structure in *Leishmania major*. Gene 385, 28–40.
[9] Sharp, P.M. and Li, W.H. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.
[10] Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.
[11] Lobry, J.R. and Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180.
[12] Das, S., Ghosh, S., Pan, A. and Dutta, C. (2005) Compositional variation in bacterial genes and proteins with potential expression level. FEBS Lett. 579, 5205–5252.
[13] Greenacre, M. (1984) Theory and Application of Correspondence Analysis, Academic, London.
[14] Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.
[15] Sueoka, N. (1992) Directional mutation pressure, selective constraints ad genetic equilibria. J. Mol. Evol. 34, 95–114.
[16] Machado, C.R., Augusto-Pinto, L., McCulloch, R. and Teixeira, S.M. (2006) DNA metabolism and genetic diversity in Trypanosomes. Mutat. Res. 612, 40–57.
[17] Dubois, M.E., Demick, K.P. and Mansfield, J.M. (2005) Trypanosomes expressing a mosaic variant surface glycoprotein coat escape early detection by the immune system. Infect. Immun. 73, 2690–2697.
[18] Buscaglia, C.A., Campo, V.A., Di Noia, J.M., Torrecilhas, A.C., De Marchi, C.R., Ferguson, M.A., Frasch, A.C. and Almeida, I.C. (2004) The surface coat of the mammal-dwelling infective trypomastigote stage of *Trypanosoma cruzi* is formed by highly diverse immunogenic mucins. J. Biol. Chem. 279, 15860–15869.
[19] Seligmann, H. (2003) Cost-minimization of amino acid usage. J. Mol. Evol. 56, 151–161.
[20] Chanda, I., Pan, A. and Dutta, C. (2005) Proteome composition in *Plasmodium falciparum*: higher usage of GC-rich nonsynonymous codons in highly expressed genes. J. Mol. Evol. 61, 513–523.

[21] Sharp, P.M. and Li, W.H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 222–230.
[22] Alvarez, F., Robello, C. and Vignali, M. (1994) Evolution of codon usage and base contents in kinetoplastid protozoans. Mol. Biol. Evol. 11, 790–802.
[23] Stevens, J.R., Noyes, H.A., Dover, G.A. and Gibson, W.C. (1999) The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*. Parasitology 118, 107–116.