BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(18)

Comparative Genomics of *Trypanosomatid* Pathogens using Codon Usage Bias

Mayank Rashmi[1] & D. Swati[1, 2]*

[1]Department of Bioinformatics, MMV, Banaras Hindu University, Varanasi-221005, India; [2]Department of Physics, MMV, Banaras Hindu University, Varanasi 221005, India; D. Swati - Email: swatid@gmail.com; *Corresponding author

Received October 15, 2013; Accepted October 19, 2013; Published November 11, 2013

Abstract:
It is well known that an amino acid can be encoded by more than one codon, called synonymous codons. The preferential use of one particular codon for coding an amino acid is referred to as codon usage bias (CUB). A quantitative analytical method, CUB and a related tool, Codon Adaptative Index have been applied to comparatively study whole genomes of a few pathogenic *Trypanosomatid* species. This quantitative attempt is of direct help in the comparison of qualitative features like mutational and translational selection. Pathogens of the *Leishmania* and *Trypanosoma* genus cause debilitating disease and suffering in human beings and animals. Of these, whole genome sequences are available for only five species. The complete coding sequences (CDS), highly expressed, essential and low expressed genes have all been studied for their CUB signature. The codon usage bias of essential genes and highly expressed genes show distribution similar to codon usage bias of all CDSs in *Trypanosomatids*. Translational selection is the dominant force selecting the preferred codon, and selection due to mutation is negligible. In contrast to an earlier study done on these pathogens, it is found in this work that CUB and CAI may be used to distinguish the *Trypanosomatid* genomes at the sub-genus level. Further, CUB may effectively be used as a signature of the species differentiation by using Principal Component Analysis (PCA).

Abbreviations: CUB: Codon Usage Bias, CAI: Codon Adaptative Index, CDS: Coding sequences, t-RNA: Transfer RNA, PCA: Principal Component Analysis.

Keywords: Codon Usage Bias, Codon Adaptation Index, Essential genes, Highly expressed genes, Translational selection, t-RNA.

Background:

Protozoan parasites from the order of *Kinetoplastida* cause diseases in humans as well as animals. Eleven genus of *Trypanosomatidae* have 565 species. Out of these, only *Leishmania* and *Trypanosoma* cause disease in humans. *Trypnosoma brucei*, *Trypanosoma cruzi* and various species of *Leishmania* are responsible for African Trypanosomiasis, Chagas disease and Leishmaniasis, devastating diseases that afflict millions of victims, largely in the tropical regions of the globe. So a comprehensive study of these genomes is expected to generate better comparative protein studies which may lead to more efficient drug designing. These single cell eukaryotes occupy a very ancient position on the evolutionary tree which diverged from the ancestor of the main eukaryotic branch **[1, 2]**. *Trypanosomatids* have unique mechanisms for gene expression, such as poly-cistronic transcription, trans-splicing, and the involvement of Pol I in the synthesis of mRNA and RNA editing **[3]**.

Comparative genomics provides a gateway to understand the evolutionary and functional relations between different species. Codon usage bias (CUB) is one of the useful analytical methods for comparative genomics. Codon usage bias is needed for the selection of translational efficiency and hence gene expression. Several steps in the gene expression process may be modulated, including transcription, RNA splicing, translation, and post-translational modification of a protein. More than one triplet of nucleotides (codon) may code an amino acid. These codons are known as synonymous codons. The unequal usage of a particular codon is termed codon usage bias. **[4]**.The analysis of CUB may aid in understanding the expression of genes and pathogenesis of any organisms and allow re-engineering of the target genes to improve their expression for gene therapy **[5]**. The Codon Adaptation Index (CAI) value is useful in the study of gene expression level of a particular gene in an organism. The value of CAI nay be used to study of the gene expression level within species, the level of heterogeneous gene expression, and comparison of the codon usage in different organisms and identification of the protein coding reading frames **[6]**.

David Horn **[7]** had studied the codon usage pattern of some tandem highly expressed genes of *Trypanosomatids*. He found the codon usage patterns of *L.major*, *T.bruei* and *T. cruzi* to be similar. Since one cannot determine the behavior of the entire genome on the basis of a particular group of genes, this study follows a more general approach. This is now possible due to more data for the coding sequences (CDS) and whole genome sequences being available.

Generally three basic questions are addressed in this research work. First-is the codon distribution of all CDSs conserved among species or not; second-is the translational selection in *Trypanosomatids* on the basis of cognate copy number of t-RNA or not; and third-is the distribution of codon usage and amino acid frequency for CDSs, highly expressed genes, low expressed genes and essential genes in *Trypanosomatids* similar or not.

It is found that the pattern of codon usage bias is more or less similar between all highly expressed, less expressed genes and complete CDS for a given genome, but the difference of CUB of different genomes is enough to distinguish the genomes at the sub- genus level. Principal Component Analysis (PCA) was used to compare the CUB of all five *Trypanosomatid* species, with *Plasmodium vivax*, an Apicomplexan genome taken as an outlier with GC% value close to that of *T. brucei*. The resulting plot clearly differentiates CUB signature of *L.braziliensis* from that of *L.major*, and *T.brucei* and *T.cruzi* are also shown well-separated as different species. However *L.major* Friedlin and *L.infantum*, the old world species, that have 92 % identity even at amino acid level **[8]** cannot be distinguished. This behavior is as expected and they may be treated as a single species.

Methodology:

We have chosen five species to work on, namely- *Leishmania major* Friedlin, *Leishmania infantum* JPCM5, *Leishmania braziliensis* MHOM, *Trypanosoma brucei brucei* 927 and *Trypanosoma cruzi* CL Brener. We have also taken *Plasmodium vivax* as as outlier genome for PCA calculation.

Materials

All the coding sequences for generating the codon usage table are obtained from NCBI (http: //www.ncbi.nlm.nih.gov/) and TriTrypDB (http://tritrypdb.org/tritrypdb/). The numbers of CDSs analyzed are 8102 in *L.major*, 7932 in *L.infantum*, 7834 in *L.braziliensis*, 8663 in *T. brucei* and 8990 CDSs of *T.cruzi* **(see Table 1).** The frequency of codons is calculated using CALCULATION OF PARAMETERS server (http://genomes.urv.cat/CAIcal). Only those sequences which are less than ten thousand nucleotides are considered because longer sequences contain repetitive regions with an in-built bias, and therefore are excluded from this study **[9]**. The DNA sequences of some factors which play a central role in replication and transcription process were downloaded from (http: //www.ncbi.nlm.nih.gov/).

Methods

t-RNA sequences of *Trypanosomatids* are obtained from GeneDB (ttp://www.genedb.org). Anti-codons are predicted by using TFAM Webserver 1.3 (http://tfam. lcb.uu.se/) for analyzing the copy number of t-RNA of the corresponding codon **[10]**.). The codon adaptation index (CAI) is used to predict gene expression levels and their value calculated by CAI CAL (CAI) (http://genomes.urv. cat/CAIcal/) for these factors. Codon frequency for highly expressed genes and low expressed genes are calculated by CODONW 1.3 (John Peden and fttp://molbiol.ox.ac.uk/cu/codonW.tar.Z/).

PCA is a useful statistical technique that finds patterns in data of high dimensionality. It reduces the variables by using suitable coordinate transformations without losing relevant information. Codon usage of sixty one codons (excluding the three stop codons) of the above five species and *P.vivax* have been taken as input variables in PCA analysis by SPSS 16.0.

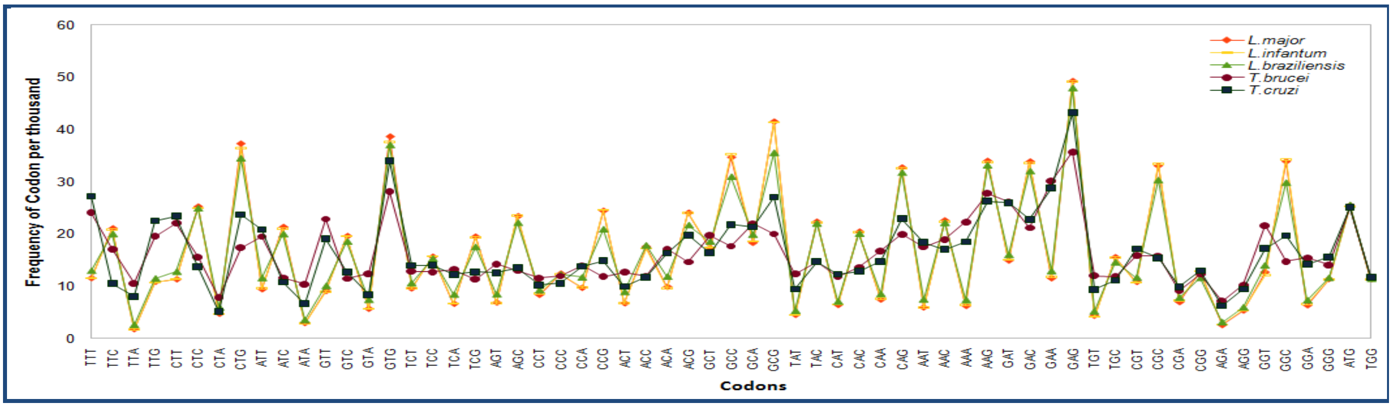

Figure 1: Plot of frequency of codon usage per thousand in all CDSs of length of less than ten thousands in *L.major* Freidlin(Orange), *L.infantum* JPCM5(Yellow), *L.braziliensis* MHOM(Green), *T.brucei* 927(Brown) and *T.cruzi* CL Brener (Blue).

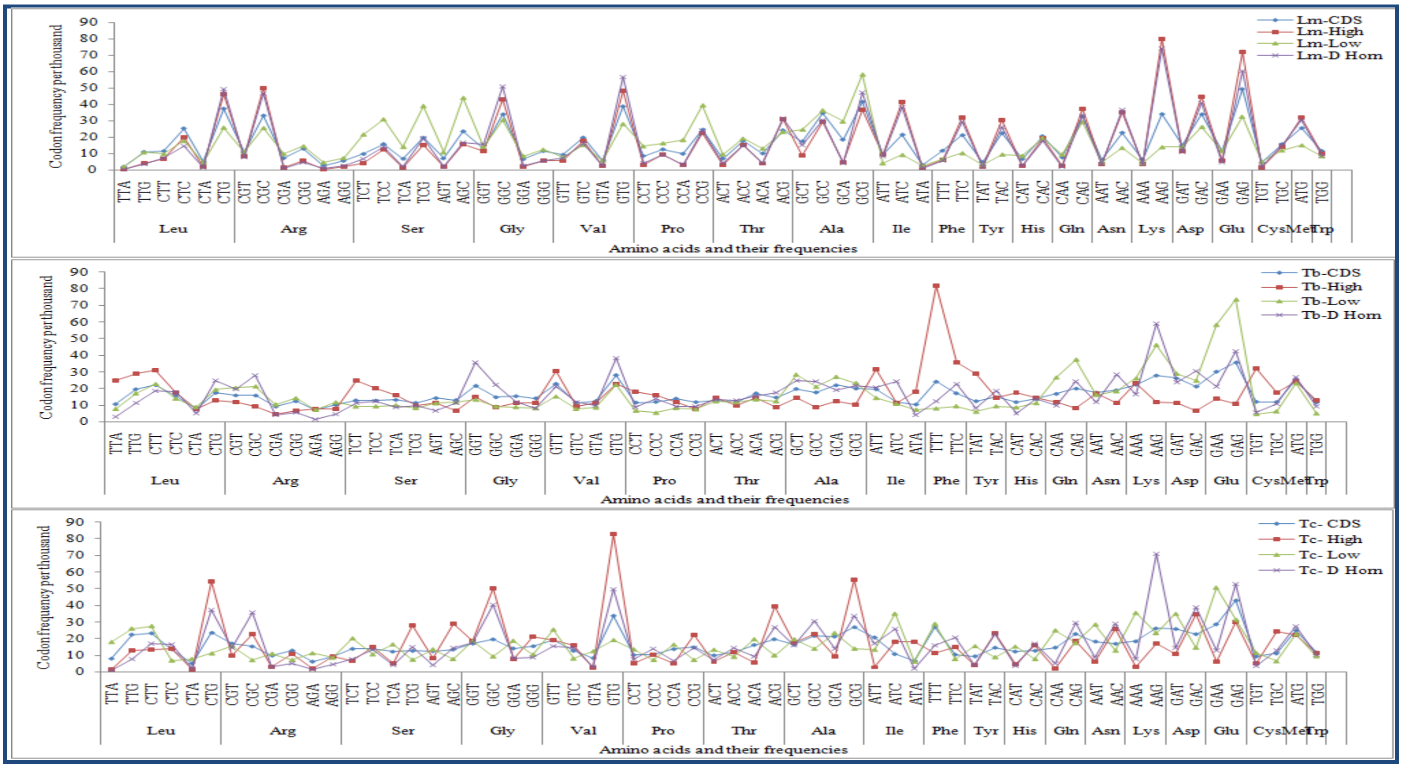

Figure 2: Plot of frequency of codon usage per thousand for different types of genes in *Leishmania major*(Lm), *Trypanosoma brucei*(Tb) and *Trypanosoma cruzi*(Tc).

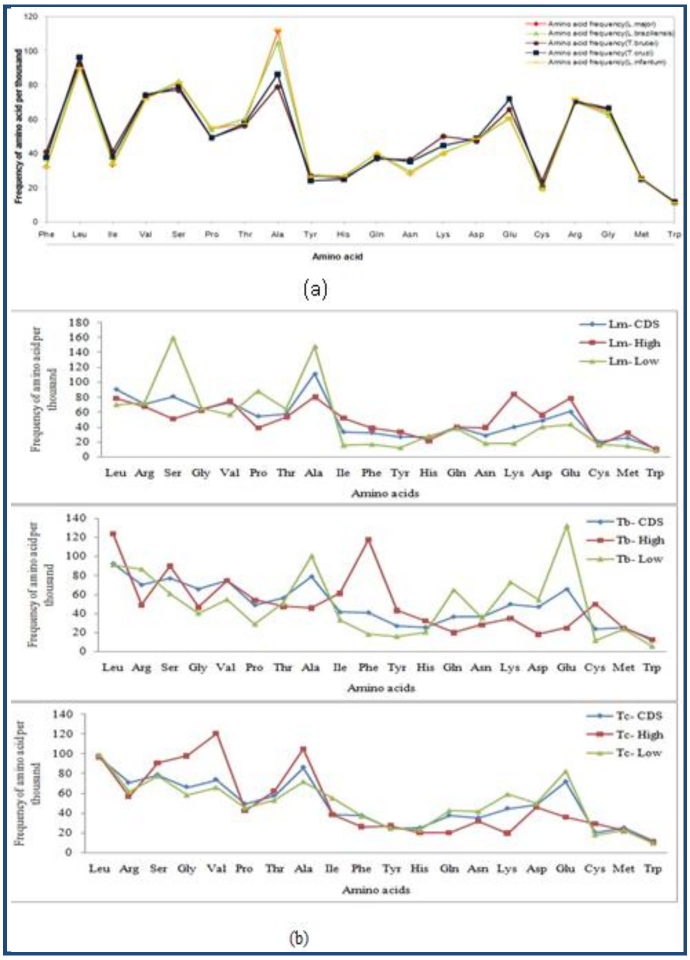

Figure 3: a) Plot of frequency of amino acids versus amino acids in *L.major* Freidlin(Orange), *L.infantum* JPCM5(Yellow), *L.braziliensis* MHOM(Green), *T.brucei* 927(Brown) and *T.cruzi* CL Brener (Blue); **b)** Plot of frequency of amino acid per thousand for different types of genes in *Leishmania major*(Lm), *Trypanosoma brucei*(Tb) and *Trypanosoma cruzi*(Tc).

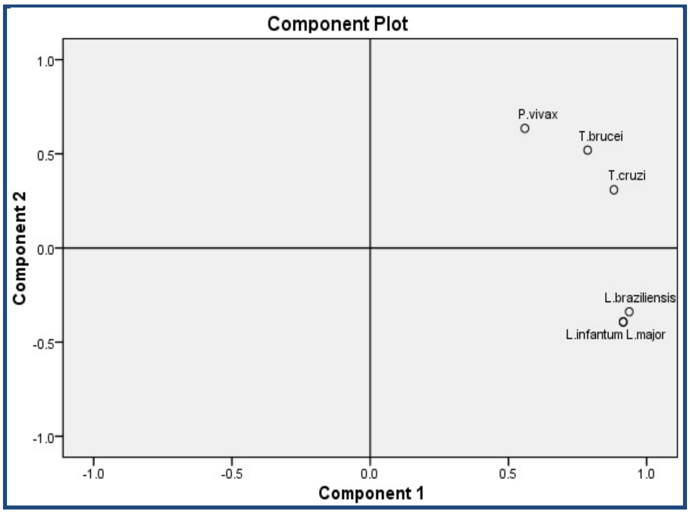

Figure 4: PCA plot for six species by using codon frequency per thousand.

Results & Discussion:

Codon usage bias analysis

The genome size of *Trypanosomatids* is very large, average genome size in four species is 30 Mb but *T.cruzi* has a larger size of about 90 Mb **Table 1 (see supplementary table)** because 50% of its genome contains repetitive elements **[11]**. Overall codon frequencies of *Trypanosomatids* are listed in **Table 2 (see supplementary table)**. The dominant codons referred to in **Table 2** are those that are at least 1.25 times more frequent than the expected frequency of synonymous codons that code a specific amino acid. High codon usage bias in *Leishmania* species and comparatively lesser codon usage bias in *Trypanosoma* species are seen **(Figure 1).** Bias of cognate codon of the *Trypanosoma* species are not conserved as much as in the *Leishmania* species.

Translational selection in Trypanosomatids

Transfer RNA plays a very important role in decoding the genome in protein synthesis. All amino acids are coded by the three letter code of t-RNA called an anticodon. It is usually composed of adenosine (A), guanosine (G), cytidine (C) and uridine (U) **[12]**. Specific t-RNAs are responsible for encoding the multiple codons that differ only in the third position and give rise to the wobble hypothesis **[13]**. In *Trypanosomatids* t-RNA genes are present only for 43 to 45 codons **(Table 1)**. The total copy numbers of t-RNAs show variability in the *Trypanosomatids***.** There are 66 in *L. infantum,* 66 in *L. braziliensis* and 64 in *T.brucei.* But the number of t-RNAs is 83 in *L. major* and 115 in *T.cruzi.*

Of the four nucleotides in any DNA sequence, C and T are pyrimidines and A and G are purines. There are three different types of wobbling possible of an anticodon in the process of translation. Case 1 is when wobbling occurs between pyrimidine and pyrimidine; Case 2 occurs when wobbling is between purine and purine; Case 3 involves wobbling between pyrimidine and purine. About 13-16 codon pairs follow Case 1. The preponderance of pyrimidines at the third position of a codon as observed by Wald is seen in this study also **[14].**

It is noteworthy that only three types of wobbling codon pairs are observed out of eight possiblities of wobbling between pyrimidine and purine the details of which are given in **(Table 2).**

Comparison between codon usage pattern of complete CDSs and other genes

Highly expressed genes tend to use only a limited number of codons and display a high codon bias due to translational selection **[15]**. High value of CAI of any gene represents higher expression of that gene. The value of 1.0 for CAI indicates the maximum usage of these codons and lower value shows the usage of less preferred codons **[6]**. The CAI is used to find out whether the expression of selected genes is low or high. The codon usage of some replication and transcription factors which are coded by essential genes is also analyzed. The study of essential genes is helpful in drug designing because these genes are important for the survival of organisms.

CUB largely reflects the differences in the selection pressure on synonymous codons and this selection pressure leads to a substitution in nucleotides composing the codons. This nucleotide substitution is inversely proportional to the degree of codon bias **[16]**. The rate of substitution is less in coding sequences as compared to non-coding regions. It is still lesser in highly expressed genes **[17]**. Codon usage frequencies of different types of genes are shown in **(Figure 2).** All CDS and essential genes show similar frequency of codons but highly and less expressed genes show different result in both *Leishmania* and *Trypanosoma* species. In *Leishmania* the frequencies of codons of highly expressed genes are higher than the value for overall CDSs **[6]**.

Amino acids frequencies in all CDS and other genes

Amino acid frequency is defined as the total frequency of their respective codons. It is almost the same in both the genus but that of Proline(P), Alanine(A), Asparagine(N), Lysine(L) and Glutamic acids are observed to vary. The frequencies are higher for Proline(P) and Alanine(A) in the *Leishmania* as compared to the *Trypanosoma*. But frequencies of Asparagine (N), Lysine (K) and Glutamic acid (E) in the *Trypanosoma* species are higher than in the *Leishmania* species **(Figure 3)**. The frequencies of amino acids have similar distribution in CDSs and essential genes**.**

At the level of properties, the amino acids are divided into four classes, hydrophobic, charged, polar/uncharged and special amino acids **[18].** Charged and Polar amino acids may be grouped together as hydrophilic and are shown as such in **Table 3 (see supplementary table).** Frequencies of hydrophilic amino acids are slightly higher than hydrophobic amino acids in all chosen *Trypanosomatids.*

Principal Component Analysis is a very useful method for multi-variate analysis. There are 61 variables for the six species chosen. The method uses coordinate transformations to reduce the components to a few principal ones. The first component, PC1 and the second component, PC2 have subsumed 98.5% property of whole 61 variables in every species. The role of PC1 may be deduced to be related to the GC% of the CDS since the CUB values of outlier genome of *P. vivax* lies close to that of the *T.brucei* as expected and shown in **(Figure 4),** and GC% is found to be one of the controlling parameters shaping CUB **[6]**.

Conclusions:

t-RNA abundance, shared specific mutational bias, gene expression level, amino acid composition and GC composition are different factors influencing codon usage bias**[19-23]**.The occurrence of codon usage varies from species to species. The codon usage bias is more conserved in the *Leishmania* than in the *Trypanosoma.* There is a significant variation in overall trends between *T. brucei* and *T.cruzi.* It is probably due to at least 50% of the *T. cruzi* genome being composed of repeats, consisting mostly of large gene families of surface proteins, retro-transposons, and subtelomeric repeats **[24]**. The two species of *Trypanosoma* genus *T.brucei* and *T.cruzi* belongs to two sub- genus Trypanozoon and Schizotrypanum. It is found in this study that the two sub- genus are differentiated at every level of analysis. The use of PCA identified the differences at species level.

Analysis of amino acids cannot give information about mutational selection because it may be possible that the amino acids will not change due to mutation in one nucleotide. Mutation in third position of codon cannot change the amino acid but mutation at the first and second position may change the amino acid. Because of the high degree of conservation at the amino acid level, we may conclude that for *Leishmania*, mutation occurs only at the third or synonymous position of codons. But in the *Trypanosoma* species studied here a mutation in the first and second position of the codon changes the amino acid. Therefore we can say that difference in frequency of these amino acids in *Trypanosoma* is the result of mutation of first position on that cognate codon. For example, lysine frequency is high in *T.brucei* but frequency of glutamic acid is high in *T.cruzi*. Lysine is encoded by AAA and AAG, and glutamic acid by GAA and GAG, showing that a mutation of first position of the codon can change the amino acid. It is affirmed that difference in frequency of these amino acids in *Trypanosoma* is the result of mutation of the nucleotide in the first position of that cognate codon **(Table 2).**

If the frequency of a particular codon in a genome is very high, then the copy number of the corresponding t-RNA is also high. A direct correlation in these variables is found (result not shown). In this study it is found that few codons of *Trypanosomatids* are highly preferred but their cognate t-RNA is absent. This indicates that wobbling plays a very important role in translational selection but it may also be possible that the data is incompletely represented or the database used is not comprehensive. The positive correlation between amino acid and t-RNA shows that translational selection is a major force affecting codon usage bias in *Trypanosomatids*.

The essential genes of *Leishmania* and *Trypanosoma* are homologs. But codon preference in essential genes of *Leishmania* is different from that of *Trypanosoma*. This is found to be primarily due to the difference in GC% (PCA result).The codon usage table is used as a reference table for calculation of CAI and estimating the gene expression from it. In other words, preferred codons of all essential genes and highly expressed genes are not similar in different species of *Trypanosomatids*, exception being *L.major* and *L.infantum*. Sequence conservation between *Leishmania* species is high, the average amino acid identity between *L. major* and *L.infantum* is 92%, and the average nucleotide identity is 94% **[8]**.The study here also confirms that the variation of all properties including the two principal components in the PCA for *L.major* and *L.infantum* are almost identical. For all intents and purposes they can be taken to be one genome. All CDSs, essential genes, highly expressed genes and less frequently expressed genes have been analyzed and it is clear that the result is different from the previous study of D. Horn **[7]** and his deductions. It is confirmed that the CUB or codon usage bias may be used as a tool for detecting differences at the species level, and is therefore a useful tool for comparative genomics.

Acknowledgement:
MR gratefully acknowledges the award of Junior Research Fellowships from the Council for Scientific and Industrial Research, India, which aided in conducting this study.

References:

[1] Landfear SM, *PNAS*. 2003 **100**: 7 [PMID: 12509502]
[2] Aguero F *et al. Genome Res*. 2000 **10:** 1996 [PMID: 11116094]
[3] Calvillo SM *et al. J Biomed Biotechnol*. 2010 **2010:** 525241 [PMID: 20169133]
[4] Fuglsang A, *Mol Biol Evol*. 2006 **23:** 1345 [PMID: 16679346]
[5] Lu H *et al. Acta Biochim Biophys Sin*. 2005 **37:** 1 [PMID: 15645075]
[6] Sharp PM & Li WH, *Nucleic Acids Res*. 1987 **15:** 1281 [PMID: 3547335]
[7] Horn D, *BMC Genomics*. 2008 **9:** 2 [PMID: 18173843]
[8] Peacock CS *et al. Nature Genet*. 2007 **39:** 839 [PMID: 17572675]
[9] Puigbo P *et al. Biology Direct*. 2008 **3:** 38 [PMID: 18796141]
[10] Taquist H *et al. Nucleic Acids Res*. 2007 **35:** W350 [PMID: 17591612]
[11] Arner E *et al. BMC Genomics*. 2007 **8:** 391 [PMID: 17963481]
[12] Agris PF, *Nucleic Acids Res*. 2004 **32:** 223 [PMID: 14715921]
[13] Crick FHC, *J Mol Biol*. 1966 **19:** 548 [PMID: 5969078]
[14] Wald N *et al. Nucleic Acids Res*. 2012 **40**: 7074 [PMCID: PMC3424539]
[15] Grantham R *et al. Nucleic Acids Res*. 1981 **9:** 43 [PMID: 7208352]
[16] Sharp PM & Li WH, *Nucleic Acids Res*. 1986 **14:** 7737 [PMCID: PMC311793]
[17] Li WH *et al. Nature*. 1981 **292:** 237 [PMID: 7254315]
[18] Umbarger HE, *Annu Rev Biochem*. 1978 **47**: 532 [PMID: 354503]
[19] Wan XF *et al. BMC Evol Bio*. 2004 **4:** 19 [PMID: 15222899]
[20] Sueoka N & Kawanishi Y, *Gene*. 2000 **261:** 53 [PMID: 11164037]
[21] Sueoka TK *et al. Gene*. 1999 **238:** 59 [PMID: 10570984]
[22] Black WJ *et al. Nature* 2003 **422:** 633 [PMID: 12687005]
[23] Ikemura T, *Mol Biol Evol*. 1985 **2:** 13 [PMID: 3916708]
[24] El-Sayed NM *et al. Science* 2005 **309:** 409 [PMID: 16020725]

Edited by P Kangueane
Citation: Rashmi & Swati, Bioinformation 9(18): 912-918 (2013)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(18)

Supplementary material:

Table 1: Genome properties of chosen Trypanosomatids

Name of *Trypanosomatids*	Number of Chromosomes	Genome Size (Mb)	Total number of CDS	Number of Analyzed CDS#	Number of Non analyzed CDS##	Overall G+C % in genome	G+C % in CDSs	Copy number of t-RNA gene	Number of t-RNA
L.major Friedlin	36	32.82	8167	8102	65	59.72	61.42	83	45
L.infantum JPCM5	36	32.12	7994	7932	62	59.45	61.45	66	43
L.braziliensis MHOM	35	37.28	7895	7834	61	57.59	59.55	66	44
T.brucei 927	11	26.08	8712	8663	49	46.43	50.23	64	44
T.cruzi CL Brener	41	89.61	9042	8990	52	56.93@	52.46	115	45

@-This value is very high as compared to *T.bruceli* due to presence of very large number of N (Unidentified) nucleotides at the telomeric region; #-Less than ten thousand length; ##-More than ten thousand.

Table 2: Codon frequency table and the cognate copy number of t-RNA in *Trypanosomatids*

Amino Acids (AA)	Symble of AA	Codons	*L.major* Friedlin				*L.infantum* JPCM5				*L.braziliensis* MHOM				*T.brucei* 927				*T.cruzi* CL Brener			
			Frequency per thousand	t-RNA	Case	Dominant Codon	Frequency per thousand	t-RNA	Case	Dominant Codon	Frequency per thousand	t-RNA	Case	Dominant Codon	Frequency per thousand	t-RNA	Case	Dominant Codon	Frequency per thousand	t-RNA	Case	Dominant Codon
Phe	F	TTT/TTC	11.47/21	0/2	1a	TTC	11.49/20.78	0/2	1a	TTC	12.95/19.97	0/1	1a	TTC	24.03/17.09	0/2	1a	-	27.19/10.44	0/4	1a	TTT
Leu	L	TTA/TTG	1.7/10.78	2/1	2a	CTC	1.72/10.68	1/1	2a	CTC	2.6/11.47	1/0	2a	CTC	10.55/19.53	1/0	2a	TTG	7.98/22.5	2/2	2a	TTG
		CTT/ CTC	11.24/25.17	3/0	1b	CTG	11.33/24.93	1/0	1b	CTG	12.75/24.88	2/0	1b	CTG	22.02/15.56	2/0	1b	CTT	23.35/13.66	4/0	1b	CTT
		CTA/ CTG	4.64/37.21	1/2	-		4.66/36.38	1/2	-		5.91/34.45	1/2	-		7.76/17.36	1/1	2a		5.11/23.63	2/2	2a	CTG
		CTG/CTC	37.21/25.17	2/0	3a		36.38/24.93	2/0	3a		34.45/24.88	2/0	3a		17.36/15.56	1/0	3a		23.63/13.66	2/0	3a	
Ile	I	ATT/ ATC	9.33/21.26	3/0	1b	ATC	9.53/20.97	1/0	1b	ATC	11.55/19.94	2/0	1b	ATC	19.48/11.6	2/0	1b	ATT	20.79/10.75	4/0	1b	ATT
		ATA/ATC	2.88/21.26	1/0	3c		2.89/20.97	1/0	3c		3.55/19.94	1/0	3c		10.37/11.6	1/0	3c		6.59/10.75	3/0	3c	
Val	V	GTT/ GTC	9/19.53	2/0	1b	GTG	8.9/19.53	3/0	1b	GTG	10.05/18.51	2/0	1b	GTG	22.82/11.4	2/0	1b	GTG	19.04/12.73	2/0	1b	GTG
		GTA/ GTG	5.61/38.57	1/2	-		5.65/37.52	1/2	-		7.39/36.98	1/1	2a		12.36/28.05	1/1	2a		8.33/33.91	2/4	2a	
		GTT/GTG	9/38.57	2/2	3b		8.9/37.52	3/2	3b		10.05/36.98	2/1	3b		22.82/28.05	2/1	3b		19.04/33.91	2/4	-	
		GTG/GTC	38.57/19.53	2/0	3a		37.52/19.53	2/0	3a		36.98/18.51	1/0	3a		28.05/11.4	1/0	3a		33.91/12.73	4/0	3a	
Ser	S	TCT/ TCC	9.5/15.57	1/0	1b	AGC	9.6/15.66	1/0	1b	AGC	10.57/15.26	1/0	1b	AGC	12.82/12.69	1/0	1b	-	13.92/14.01	2/0	1b	-
		TCA/ TCG	6.56/19.42	1/1	2a	TCG	6.67/19.29	1/1	2a	TCG	8.4/17.48	1/0	2a	TCG	13.13/11.32	1/1	-		12.21/12.69	2/2	-	
		TCG/TCC	19.42/15.57	1/0	3a		19.29/15.66	1/0	3a		17.48/15.26	0/0	-		11.32/12.69	1/0	3a		12.69/14.01	2/0	3a	
		AGT/ AGC	6.8/23.34	0/2	1a		6.69/23.52	0/2	1a		8.44/22.11	0/2	1a		14.2/12.96	0/2	1a		12.53/13.5	0/2	1a	
Pro	P	CCT/ CCC	8.27/12.34	2/0	1b	CCG	8.61/12.25	2/0	1b	CCG	9.25/12.24	1/0	1b	CCG	11.51/11.91	1/0	1b	-	10.17/10.55	2/0	1b	-
		CCA/ CCG	9.61/24.36	1/2	-		9.77/24.56	1/2	-		11.68/20.88	1/2	-		13.86/11.79	1/1	-		13.75/14.82	2/2	-	

© 2013 Biomedical Informatics

		CCG/CCC	24.36/12.34	2/0	3a		24.56/12.3	2/0	3a		20.88/12.24	2/0	3a		11.79/11.91	1/0	3a		14.82/10.55	2/0	3a	
Thr	T	ACT/ ACC	6.6/17.5	3/0	1b	ACG	6.73/17.19	0/0	-	ACG	8.88/17.79	0/0	-	ACG	12.66/11.86	1/0	1b	-	9.92/11.68	2/0	1b	ACG
		ACA/ ACG	9.8/24	1/2	-		9.61/24.01	1/2	-		11.83/21.66	1/2	-		17.04/14.58	1/1	-		16.29/19.74	2/2	-	
		ACG/ACC	24/17.5	2/0	3a		24.01/17.19	2/0	3a		21.66/17.79	2/0	3a		14.58/11.86	1/0	3a		19.74/11.68	2/0	3a	
Ala	A	GCT/ GCC	17.2/34.6	2/0	1b	GCG	17.35/35.26	2/0	1b	GCG GCC	18.54/30.92	2/0	1b	GCG	19.71/17.62	2/0	1b	-	16.35/21.75	2/0	1b	-
		GCA/ GCG	18.23/41.42	1/2	-		18.56/41.31	1/0	2a		19.83/35.5	1/1	2a		21.92/19.94	1/2	2b		21.33/27	1/2	-	
		GCT/GCG	17.2/41.42	2/2	3b		17.35/41.31	2/0	3b		18.54/35.5	2/1	3b		19.71/19.94	2/2	-		16.35/27	2/2	3b	
Tyr	Y	TAT/ TAC	4.44/22.27	0/3	1a	TAC	4.43/22.15	0/1	1a	TAC	5.3/21.97	0/1	1a	TAC	12.34/14.64	0/2	1a	-	9.4/14.65	0/2	1a	-
His	H	CAT/ CAC	6.32/20.37	0/2	1a	CAC	6.46/20.27	0/2	1a	CAC	7.07/20.04	1/1	1b	CAC	11.79/13.61	0/1	1a	-	12.24/12.79	0/4	1a	-
Gln	Q	CAA/ CAG	7.36/32.64	1/3	-	CAG	7.56/32.49	1/3	-	CAG	8.52/31.7	1/2	2a	CAG	16.71/19.93	1/2	-	-	14.7/22.9	1/4	2b	-
Asn	N	AAT/ AAC	5.86/22.52	0/3	1a	AAC	5.91/22.35	0/1	2a	AAC	7.45/22.1	0/3	2a	AAC	17.5/18.91	0/2	2a	-	18.33/16.99	0/4	2a	-
Lys	K	AAA/ AAG	6.14/33.91	1/3	-	AAG	6.44/33.75	1/3	-	AAG	7.37/33.1	1/3	-	AAG	22.25/27.75	1/3	2b	-	18.46/26.23	2/4	-	-
Asp	D	GAT/ GAC	14.82/33.79	0/3	1a	GAC	15/33.59	0/2	1a	GAC	15.98/31.99	0/4	1a	GAC	26.1/21.15	0/2	1a	-	25.94/22.76	0/2	1a	-
Glu	E	GAA/ GAG	11.43/49.18	1/2	2a	GAG	11.78/49.11	1/1	2a	GAC	12.87/47.87	1/2	2a	GAC	30.07/35.64	1/2	2b	-	28.73/43.15	2/4	-	-
Cys	C	TGT/ TGC	4.28/15.48	0/1	1a	TGC	4.1/15.46	0/1	1a	TGC	5.17/14.55	0/1	1a	TGC	11.97/11.82	0/2	1a	-	9.33/11.11	0/2	1a	-
Arg	R	CGT/ CGC	10.74/32.98	4/0	1b	CGC	10.64/33.33	4/0	1b	CGC	11.63/30.28	2/1	1b	CGC	15.81/15.75	3/0	1b	CGT CGC	17.1/15.37	4/0	1b	CGT CGC
		CGA/ CGG	6.89/12.74	1/1	2a		7.25/12.83	1/1	2a		7.81/11.58	1/1	2a		9.08/12.34	1/1	2a		9.84/12.93	2/2	-	
		AGA/ AGG	2.49/5.27	1/1	-		2.64/5.33	1/1	-		3.12/5.97	1/1	-		7.1/10.21	1/1	-		6.22/9.4	2/2	-	
Gly	G	GGT/ GGC	12.6/33.79	0/4	1b	GGC	12.13/34.25	0/2	1b	GGC	13.95/29.78	0/1	1b	GGC	21.59/14.69	0/3	1b	GCT	17.15/19.65	0/4	1b	-
		GGA/ GGG	6.27/11.3	1/1	-		6.51/11.26	1/1	-		7.29/11.57	1/1	-		15.36/14.06	1/1	-		14.15/15.53	2/2	-	
Met	M	ATG	25.36	4	-		25.22	4	-		25.53	5	-		25.05	3	-		25.03	6	-	
Trp	W	TGG	11.32	1	-		11.22	1	-		11.41	1	-		11.57	1	-		11.7	2	-	

1a: T ending codon encoded by the t-RNA of C ending codon (pyrimidine - pyrimidine); 1b: C ending codon encoded by the t-RNA of T ending codon (pyrimidine - pyrimidine); 2a: G ending codon encoded by the t-RNA of A ending codon (purine - purine); 2b: A ending codon encoded by the t-RNA of G ending codon (purine - purine); 3a: C ending codon encoded by the t-RNA of G ending codon (purine – pyrimidine); 3b: G ending codon encoded by the t-RNA of T ending codon (pyrimidine – purine); 3c: C ending codon encoded by the t-RNA of A ending codon (purine – pyrimidine

Table 3: Table of amino acid frequency on the basis of their properties

	Amino Acid Properties	***L.major*** **Friedlin**	***L.infantum*** **JPCM5**	***L.braziliensis*** **MHOM**	***T.brucei*** **927**	***T.cruzi*** **CL Brenner**
For all CDSs	Hydrophobic	404.25	402.46	401.95	392.77	393.22
	Hydrophilic	454.59	455.7	458.87	464.96	470.56
	Special	138.3	138.9	136.36	138.56	136.22
For essential Genes	Hydrophobic	420.35	404.5	401.1	382.4	394.62
	Hydrophilic	446.67	448.0	461.79	471.41	472.29
	Special	142.18	145.08	135.94	139.75	131.13

Hydrophobic – LMIVWAFY; Hydrophilic – DEKRHSTNQ; Special - GCP