Parasitology International 58 (2009) 215–219

Contents lists available at ScienceDirect

Parasitology International

journal homepage: www.elsevier.com/locate/parint

The composition of untranslated regions in *Trypanosoma cruzi* genes

Adeilton Brandão [a,*], Taijiao Jiang [b]

[a] Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil
[b] Center for Computational and Systems Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 9 May 2009
Received in revised form 26 May 2009
Accepted 1 June 2009
Available online 6 June 2009

Keywords:
Trypanosoma cruzi
UTR
Untranslated region
Trinucleotide
Transcription
SSR
Simple Sequence Repeat

ABSTRACT

We collected the UTRs from *Trypanosoma cruzi* genes that have been experimentally mapped and are publicly available, and made a comprehensive analysis of their composition features including sequence length, G+C content and relationship to ORF, composition of the most frequent words, and distribution of Simple Sequence Repeats (SSR). *T. cruzi* UTRs exhibit range length of 10–400 bp for 5′ UTR and 17–2800 for 3′ UTR. Both UTRs display mean G+C content of 40%. Ratios between the UTR and protein coding segments show that the 5′ UTR is limited to a maximum of 20% of the total length in the final transcript. The 5′ UTR most frequent words in the range 4–12 bases are almost exact complement to the 3′ UTR respective words. SSR in 3′ UTR are longer than in 5′ UTR and are mostly derived from TA/AT, TG/GT, and TTA/ATT. SSR accounts up to 20% of the nucleotide composition in 5′ UTR and up to 90% in the 3′ UTR.

1. Introduction

Trypanosoma cruzi has both a genome organization and transcription pattern that differ to a large extent from the ones observed in higher eukaryotes, e.g., majority of genes with no introns, polycistronic transcription and mRNA processing by trans-splicing [1]. Gene description in *T. cruzi* has not been followed by the detailed analysis of constituent elements such as the messenger RNA (mRNA) leader segment, also known as the Untranslated Region (UTR). This absence of UTR annotation in genomes is not exclusively to *T. cruzi*: most gene prediction algorithms ignore the UTRs in their searching for coding content [2]. UTRs exert their activities through certain components that range from simple DNA elements (e.g., sequence repeats) to highly ordered secondary structures embedded in the 5′ or 3′ UTR as well as the presence of small open reading frames (uORFs). UTRs may control translation, mRNA half-life, interactions to proteins, and communication among UTR themselves [3]. In *T. cruzi*, both UTRs (5′ and 3′) of some genes have been shown to carry out activities at the post-transcriptional level [4–9]. Despite these experimental findings, little is known about UTR in *T. cruzi* genes. Previous work [10] commented on the need for a better knowledge of the UTR in *T. cruzi* under a population perspective, with implications in problems commonly associated with strain typing and Chagas' disease evolution. To fill this gap, we need a combination of experimental data generation and computational sequence analysis. Recently, a survey of *T. cruzi* UTRs has been carried out, but it is solely based on Expressed Sequence Tags (EST) and the information was limited to their length distribution and composition analysis of polypyrimidine tracts [11]. Here we will expand this knowledge-base by presenting a comprehensive analysis of *T. cruzi* UTRs exclusively with genes that have been experimentally mapped, i.e., sequences that were not determined by the genome project. We evaluate whether UTRs from *T. cruzi* genes have a distinct feature considering both its position in the Tree of Life [12] and the current knowledge of UTRs in multicellular eukaryotes [13]. To this end, we examined the characteristics of UTR sequences with regard to the following aspects: a) typical values of 5′ and 3′ UTR sequence lengths; b) G+C content between UTR; c) correlation between UTR lengths and the respective full transcript; d) the most frequent words for each UTR; e) composition of Simple Sequence Repeats (SSR).

2. Methods

All *T. cruzi* gene sequences used in this work were extracted from GenBank-NCBI (release 168.0, October 2008). Genes inferred from the *T. cruzi* genome project were not included because they have not been annotated with respect to UTRs. For the purpose of this work, we assembled two sets of sequences from all *T. cruzi* genes that have been experimentally mapped. The first set comprises all defined UTRs extracted from both partial and full length mRNA sequences. The second one is composed solely of full length mRNA sequences with defined UTRs at each end (e.g., the entire transcript containing 5′ UTR–ORF–3′ UTR). This full length sequence set will be used in the comparisons between UTRs and the ORF with respect to length and

* Corresponding author. Instituto Oswaldo Cruz, Fiocruz, Av. Brasil, 4.365, 21045-900 Rio de Janeiro-RJ, Brazil. Tel.: +55 21 2562 1396.
E-mail address: abrandao@fiocruz.br (A. Brandão).

1383-5769/$ – see front matter
doi:10.1016/j.parint.2009.06.001

Table 1
Statistics of UTR sequence lengths in experimentally mapped *T. cruzi* genes ($n = 173$ and 139, respectively for 5′ UTR and 3′ UTR).

Length (bp)	5′ UTR	3′ UTR
Minimum	11	17
Maximum	526	2847
Range	524	2830
Mean	81	334
Median	55	289
Lower quartile	36	135
Upper quartile	104	419
20th percentile	33	108
80th percentile	118	453

bp = base pairs.

composition. Sequences were manually edited to trim off segments other than UTRs and converted to fasta format. The criteria used to select the gene sequences and their respective UTRs were as follows: a) all coding sequences with annotated start codon and showing a complete 5′ UTR; b) all coding sequences with annotated termination codon and showing a complete 3′ UTR; c) all complete and annotated coding sequences showing both UTRs; d) in the case of a few sequences for which the UTRs in the mRNA have not been defined, the 5′ UTR was determined based on comparison to homologous EST sequence primed with fragment of the trans-spliced leader (the mini-exon). Similarly, The 3′ UTR was defined if a homologous EST having poly-A tail was present in the DNA databases. The concept and methods used in this analysis have been presented and reviewed in several reports [14–17]. Here, for these UTR sequences, we used online computational tools that have been developed based on these methods. They are indicated below:

1) Emboss (http://bioweb2.pasteur.fr/nucleic/intro-fr.html#prop): a file containing either all fasta formatted 5′ UTR or 3′ UTR was uploaded to the server and the following softwares were called on:

a) compseq — (word frequency 4 to 12 bases) word size = 4 up to 12; expected frequency calculated from sequence; words counted in the forward sense only.

b) geecee — (calculates fractional GC content of nucleic acid sequences) — specific parameters: none required.

2) MREPS (http://bioinfo.lifl.fr/mreps/mreps.php): simple sequence repeats search — files containing either all fasta formatted 5′ UTR or 3′ UTR sequences were uploaded to the server and the SSR searches were called on under the parameters: Resolution: 0, Allowsmall; yes, Minimal size: 4, Minimal period: 1.

3) NCBI (www.ncbi.nlm.nih.gov): BLAST homology searches, sequence retrieval, EST/cDNA comparison to genomic sequences. Searches were done either with single or batch files using the default parameters.

4) GeneDB (www.genedb.org): searches by gene names or key words in the products list of *T. cruzi* specific database were performed to get information on gene copy number.

All the output files were edited in Openoffice Calc to consolidate the new data about each UTR set. Then, we analyzed the full transcript set with respect to the association between 5′ UTR and 3′ UTR, the percent of each UTR in this final transcript, and ratios of open reading frame (ORF) to UTR length. All the statistics and graphics related to these analyses were carried out with statistical software PAST [18].

PAST functions used: 1) histogram plotting — the cumulative distribution of UTR length; bins: 10; 2) box plot — G+C content, UTR content in the final transcript and SSR content in UTR. No outliers. 3) Linear fitting with standard regression (Pearson r, p(uncorr) = probability that the values are not correlated) — function used in UTR length correlations, UTR vs ORF length correlation.

Nucleotide composition in the third codon position of ORFs was extracted with MEGA software v4 [19].

3. Results

The number of complete UTR sequences is 173 for 5′ UTR and 139 for 3′ UTR. Eighty two full length transcripts, i.e., 5′ UTR–ORF–3′ UTR were gathered from these sequences (Table 1 and Supplementary Tables S1–S3, sequences from experimentally mapped *T. cruzi* genes deposited in GenBank-NCBI up to October 2008). The typical values for *T. cruzi* UTRs are mean length of 81 bp (base pairs) for 5′ UTR and 334 bp for 3′ UTR, and median of 55 and 281 bp, respectively. In 5′ UTR 80% of length entries is below 120 bp. The 5′ UTR length distribution seems to follows the 80:20 rule, that is, 80% of the 5′ UTR lengths is in a range that correspond to 20% of the maximum observed length for a 5′ UTR in these set. The 3′ UTR sequences shows a wider range in length distribution, varying from 40 to >1000 bp. In the 3′ UTR, 80% of length is below 450 bp. It does not follow a 80:20 rule. Taking the mean values for each UTR, the ratio 3′ UTR:5′ UTR gives a value of 4.12. From the set

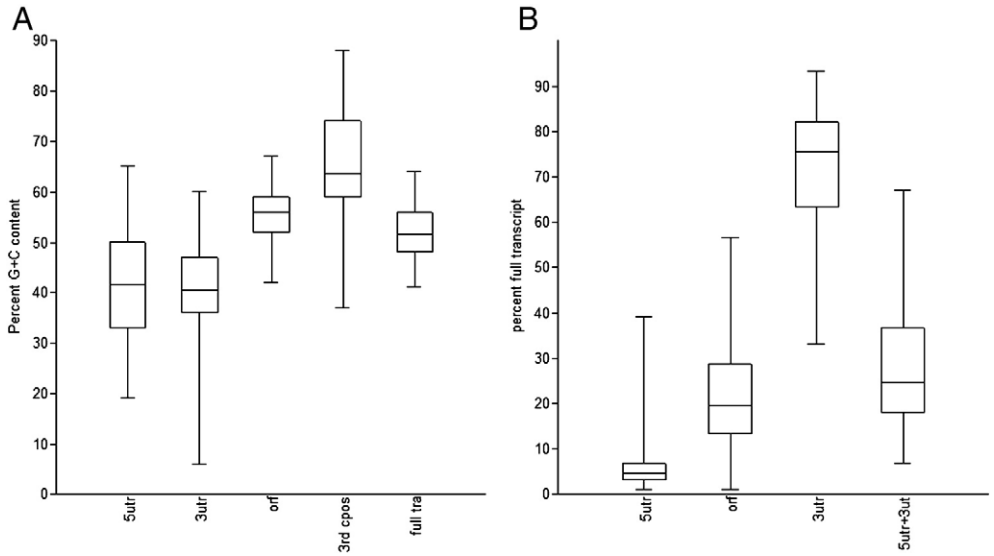

Fig. 1. G+C content and relative proportion of each mRNA component in the full length transcript (processed mRNA). A) The percent G+C content from 5′ UTR, 3′ UTR, joint UTRs (5utr + 3utr), ORF and the third codon position was plotted for the experimentally analyzed *T. cruzi* genes full transcript set. ORF and 3rd codon position exhibit higher percent G+C content than the UTR, but wide ranges are observed in UTRs and third codon position. B) The percent content of 5′ UTR, ORF, 3′ UTR and the UTR sum in the 82 *T. cruzi* full length transcripts is displayed in the respective box plot. The percent content in the final transcript appears as 5′ UTR<3′ UTR<ORF.

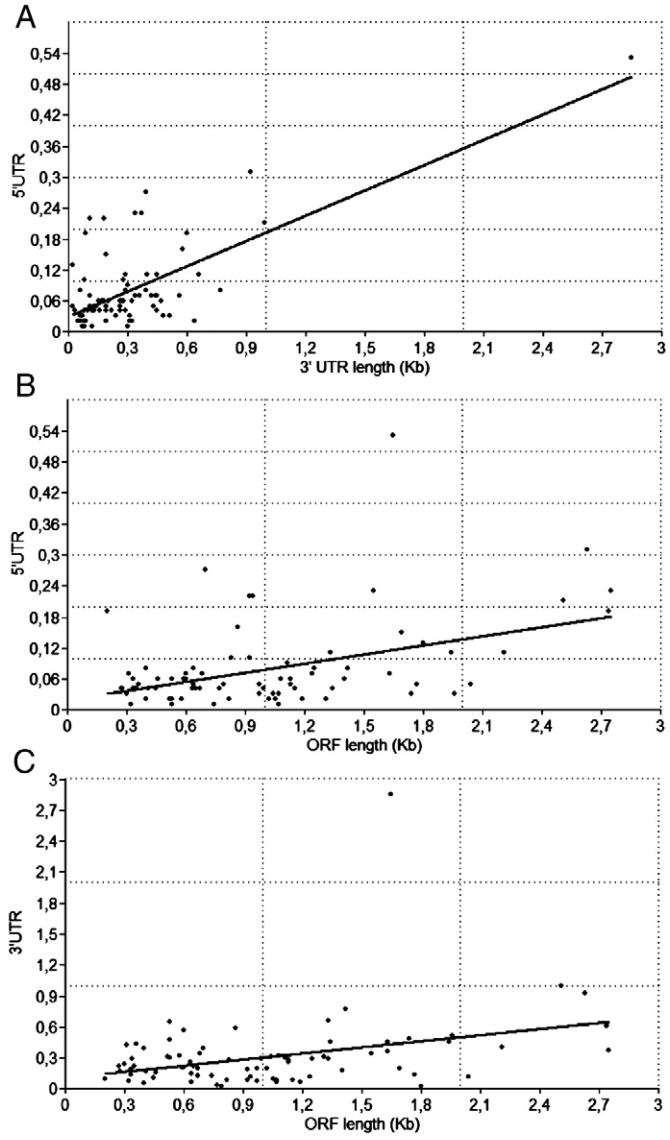

Fig. 2. Linear plot of the UTR and ORF length in the full length transcript. (A) The 5′ and 3′ UTR lengths in *T. cruzi* transcripts exhibit some linearity, as demonstrated by the high correlation coefficient ($r = 0.69$). The ORF length of the 82 full transcripts is plotted against the respective sequence length of 5′ UTR (B) and 3′ UTR (C). Weak association is observed between the lengths of each UTR and the respective ORF ($r < 0.4$). Length unit is kb.

of full length transcripts we computed the G+C content separately for 5′ UTR, 3′ UTR, ORF, the 3rd codon position and the combined sequences of both UTR (the sum of sequences 3′ UTR + 5′ UTR in the transcripts). The result is presented in Fig. 1A. The mean G+C content for 5′ UTR, 3′ UTR, ORF and third codon position is 41.10, 40.39, 55.87 and 65.71%, respectively. It is clear from these data that the protein coding segment in these full transcripts has higher G+C content and narrower range distribution than UTRs. The third codon position also has higher G+C content than UTR but displays wide range distribution. G+C content of both UTRs in this set of gene transcripts are not significantly correlated with the G+C content of the ORF or the third codon position (5′ UTR vs ORF $r = 0.1$ $p(\text{uncorr}) = 0.36$; 3′ UTR vs ORF $r = 0.28$ $p(\text{uncorr}) = 0.01$; 5′ UTR vs 3rd cpos $r = 0.13$ $p(\text{uncorr}) = 0.23$; 3′ UTR vs 3rd cpos $r = 0.28$ $p(\text{uncorr}) = 0.01$). The G+C content between UTRs is not linear. The distribution of G+C content in the full transcript approaches a normal distribution ($p(\text{normal}) = 0.616$, Shapiro–Wilk one sample normality test). An interesting observation for this full transcript set is that the G+C content of the UTRs altogether is always lower than the respective ORF and third codon position, no matter the value range each segment displays in the final transcript. For 3′ UTR in separate, only two genes (ferredoxin and metacyclin II) display the G+C content greater than the respective ORF. In contrast, the 5′ UTR G+C content is greater than the respective ORF in several genes. We can represent the segments of a transcript (the 5′ UTR, the ORF and the 3′ UTR) as percent content of the mature mRNA. Box plot of Fig. 1B summarizes the distribution for the 82 full length transcripts. As expected, the percent content in the final transcript appear as 5′ UTR<3′ UTR<ORF but UTR percent content varies from gene to gene. In this set of full transcripts, increases in 5′ UTR content do not imply a corresponding 3′ UTR content increase ($r = 0.09$ $p(\text{uncorr}) = 0.41$). In other words, less ORF percent content implies a higher 3′ UTR content regardless of the variation in 5′ UTR length. Implicitly, the total percent content of UTR in the *T. cruzi* processed mRNA is associated to the length of 3′ UTR. In general, the typical transcript in *T. cruzi* displays a maximum limit in UTR percent content of 56.65% for 3′ UTR and 39.17% for 5′ UTR. We also verified whether there are associations between lengths of UTR and ORF from these 82 transcripts. Fig. 2 shows the linear plot among the lengths of these segments. UTR lengths exhibit high linearity between themselves (correlation coefficient $r = 0.69$, $p(\text{uncorr}) < 0.001$, Fig. 2A) but to the ORF length only a weak association is observed (correlation coefficient 5′ UTR vs ORF $r = 0.44$, $p(\text{uncorr}) < 0.001$; 3′ UTR vs ORF $r = 0.34$, $p(\text{uncorr}) = 0.001$, Fig. 2B and C). Table 2 lists the most frequent words of length 4 to 12 bases in each UTR set. Each UTR is

Table 2
The three most frequent words in UTRs on the range 4–12 bases.

Word length	3′ UTR	5′ UTR
4	TTTT, AAAA, TGTG	AAAA, TTTT, CACA
5	TTTTT, AAAAA, TGTGT	AAAAA, TTTTT, CACAC
6	TTTTTT, AAAAAA, TGTGTG	AAAAAA, CACACA, TTTTTT
7	TTTTTTT, AAAAAAA, TGTGTG3T	AAAAAAA, CACACA C, TTTTTTT
8	TTTTTTTT, AAAAAAAA, TGTGTGTG	AAAAAAAA, CACACACA, TTTTTTTT
9	TTTTTTTTT, AAAAAAAAA, TGTGTGTG4T	AAAAAAAAA, TTTTTTTTT, CACACACAC
10	TTTTTTTTTT, AAAAAAAAAA, TGTGTGTGTG	AAAAAAAAAA, TTTTTTTTTT, CACACACACA
11	TTTTTTTTTT, TGTGTGTGTG T, AAAAAAAAAAA	AAAAAAAAAAA, TTTTTTTTTTT, AAGAAGAAG AA
12	TTTTTTTTTTTT, TGTGTGTGTGTG, AAAAAAAAAAAA	AAAAAAAAAAAA, TTTTTTTTTTTT, AGAGAGAGAGAG

The most frequent word in UTRs is exactly complementary to each other: (T)4–(T)12 for 3′ UTR and (A)4–(A)12 for 5′ UTR. For the second and third most frequent words, complementarity occurs only for lengths 4, 5 and 9 bases.

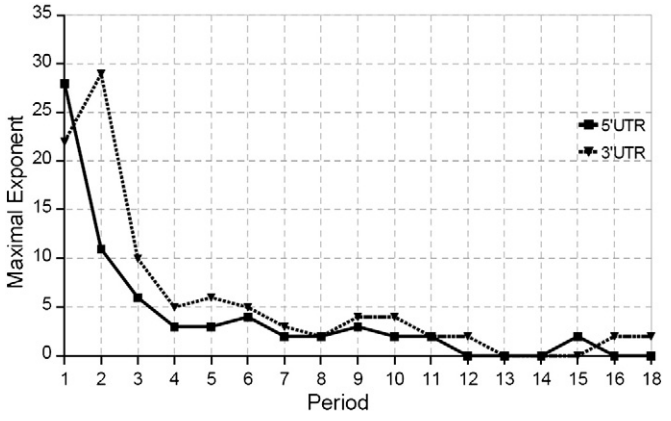

Fig. 3. Period vs maximal exponent of SSR in *T. cruzi* UTRs. The period (number of nucleotides in the repeat: 1, 2, 3, 4 and so on) and the maximal exponent (the highest count a period exhibits in a repeat) were extracted from SSR with at least 4 bases size in each UTR (mreps software). The 3′ UTR displays the highest exponents for all periods except the period one.

almost exactly the reverse of the other regarding the composition of its respective most frequent word. In 3′ UTR all most frequent words ranging from 4 to 12 are composed of nucleotide thymine. In 5′ UTR the most frequent words in this range are composed exactly by the complementary nucleotide, adenine.

SSR (Simple Sequence Repeats) are composed mainly by mono, di, tri, and tetranucleotides. Three parameters are used by most softwares to detect and count them in nucleotide sequences: period, exponent and size. For example, in the simple repeat TGTGTGTG the period is 2 (TG), the exponent is 4 (TG is repeated four times) and repeat size is 8 bases. Using the definitions of simple and maximal repeats of SSR searching software, *mreps* [16], and additionally defining a minimum size of 4 bases for periods 1 and 2, we extracted all the SSR in each UTR set. Fig. 3 displays the distribution of the counted SSR and shows a plot for periods 1 up to 18 (the highest observed in this set of UTRs) vs the respective maximum exponents. The 3′ UTR, being larger in length than the 5′ UTR, has the largest SSR both in period and exponent, except for the A mononucleotide SSR. The largest SSR dinucleotide in this UTR set is (TG)58, observed in 3′ UTR (Table 3). Dinucleotides forming the longest repeats in 3′ UTR have also the same composition in the 5′ UTR (AA, AG/GA, GT/TG, TT, AT/TA). Repeats formed by CT, TC and GG do not exhibit large sizes in 3′ UTR and in the 5′ UTR they are present at low exponents. Repeats formed by dinucleotides CA and AC also display short size in 3′ UTR, being CA the only dinucleotide in 3′ UTR whose maximum exponent is less than the corresponding one in 5′ UTR. It is also noteworthy that doublets CG/GC have similar repeat exponents in both UTRs and are the only dinucleotides forming repeats with exponents not greater than 4. The maximum values for both exponents and size in the SSR show a different distribution in each UTR (Table 3). The highest exponent in 3′ UTR occurs for a dinucleotide (TG, exponent 58) and for a mononucleotide in 5′ UTR (A, exponent 28). With respect to trinucleotide derived SSR, the observed maximum exponent is 10 in 3′ UTR (triplets ATT/TTA/TAT) and 6 in the 5′ UTR (triplet GAA). Most of the triplet SSR in 5′ UTR exhibit exponents ranging from 2 to 4 whereas in the 3′ UTR exponents range from 4 to 6. Tetranucleotide repeats are limited to exponent 3 in 5′ UTR and to 6 in 3′ UTR. In 5′ UTR SSR sequences, the less frequent dinucleotides are CC, TC, CT, and CG and the most frequent ones are AA, CA/AC, TT, and GA/AG (62.14% of SSR composition). The triplet composition in 5′ UTR SSR exhibits ATG, GAT, and TGA (0.03%, 0.04%, and 0.06%, respectively) as the lowest frequent . It is interesting that one of the stop codons (TGA) is avoided in the composition of 5′ UTR SSR. As an indication of bias, 34 triplets account for only 10% of the 5′ UTR SSR composition. At the other extreme are 10 triplets (AAA, ACA/CAC, TTT, AGA/GAG, ATA/TAT, and GTG/TGT) whose contribution to the total composition of SSR amounts to 60%. For 3′ UTR SSR the most frequent dinucleotides (TT, AA, GT/TG, AT/TA, and GG) account for 55% of total composition. The less frequent ones are CG/GC, CA/AC, and CC which account for 15% of repeat composition.

At the trinucleotide level, triplets TTT, AAA, TGT/GTG, TAT/ATA, AGA/GAG, TCT/CTC and GGG account for 61% of the SSR composition. Like the 5′ UTR SSR, 33 nucleotides account for only 10% of SSR composition. Not surprisingly, the lowest frequent trinucleotide in SSR is the stop codon TAG (0.004%). There are 3 trinucleotides pairs composed solely by GC combinations (CCG/GCC, CGC/GCG, and CGG/GGC). The contribution of these pairs to SSR composition in 3′ UTR and 5′ UTR is roughly the same: 5.7% vs 5.3%, respectively. This SSR analysis reveals that in both UTRs their main compositional features are determined by only 16% of all 64 trinucleotides.

If we sum the sizes of all SSR in 5′ and 3′ UTR individual sequences and then divide by respective UTR sequence length, we obtain a distribution of the overall SSR content in UTRs. Individually, the 3′ UTR sequences have higher SSR content and exhibit large variations, ranging from 25 to 91% of the sequence length. In 5′ UTR sequences the repeat content is lower, ranging from zero (no SSR) up to 20%.

Table 3
The largest SSRs and the highest period in UTRs.

SSR type	5′ UTR	3′ UTR
Mononucleotide	(A)28	(T)22
Dinucleotide	(GT)11	(TG)29
Trinucleotide	(GAA)6	(ATT)10
Tetranucleotide	(GGAG)3, (ATAC)3, (ACGC)3, (CGTG)3, (TCAT)3, (GCCG)3	(TAAT)6
Highest period	15, (A)6 T (A)7 T)2	18, (AC (TAT)5 T)2

4. Discussion

It is possible that our analysis is affected by the research bias in *T. cruzi* genes, which are selected for investigation based on their presumed immunological, diagnostic or therapeutic potential applications. A length distribution similar to the 5′ UTRs in higher eukaryotes is observed for this *T. cruzi* UTR set, though with a lower mean length. Some genes have their 3′ UTR length increased by the insertion of the SIRE sequences [20]. Unexpectedly, *T. cruzi* UTR mean length ratio (3′ UTR/5′ UTR = ~4) is higher than that described for plant, invertebrates and fungi (~2) [13]. This ratio is in the same range of the vertebrate UTRs, though *T. cruzi* 3′ UTRs are shorter in length. As the vertebrate 3′ UTRs are involved in several mechanisms at the post-transcriptional level and *T. cruzi* also controls gene expression at this level as well, we may expect, based on this ratio, that its 3′ UTR is extensively involved in such a control. In fact, many experimental findings including other 9 trypanosomatid species (*T. brucei* and *Leishmania* spp.) demonstrated the action of 3′ UTR in controlling the expression of important genes [21–23]. In contrast to eukaryotes from other taxonomic groups [3,13] two important differences is observed: 1) *T. cruzi* 5′ UTR mean G+C content does not differ from the 3′ UTR; 2) no significant correlation was observed between UTR G+C content and third codon position, as well as between UTRs themselves. A low or higher content of UTR is, in principle, a clue to their relative importance in mRNA metabolism. It has been observed that genes with short length UTRs are associated to higher expression [24]. Though the UTR percent content in the final transcript should be limited by the encoded protein characteristics, a peculiar feature of trypanosomatid transcription can alter this relationship regardless of ORF. That is the case of additional trans-splicing sites, which can modify the length of the 5′ UTR. Several reported genes have been demonstrated to display additional trans-splicing sites, and a recent survey of *T. cruzi* ESTs showed that this can be quite common for genes overall [25]. If an average content of UTR in *T. cruzi* mRNA is expected to have 27% of total transcript, genes that deviate too much from this value are a good experimental starting point for the demonstration of post-transcriptional mechanisms that are dependent on specific UTR motifs. Worth of note in UTR analysis is the population aspects of the genes they are derived from. In *T. cruzi* this question is of utmost importance, due to phylogenetic divergence of their populations — two major lineages evidenced by ribosomal DNA and mini-exon gene markers [26]. According to this proposed phylogeny, we should expect divergences at both functional and sequence levels for some parts of the genome. At least for the nucleotide sequence level, the 3′ UTR of an essential evolutionary conserved gene has been shown to carry on mutations according to pattern proposed by this phylogenetic model [27]. However, since UTRs are submitted to some functional pressure, we would not expect that every gene in *T. cruzi* might exhibit mutations in the non translated segments specific for this phylogenetically divergent groups. In contrast to multicellular eukaryotes [28] the 5′ UTR SSR composition does not divergemuch more than that exhibited by 3′ UTR. However, when we take into account the relative proportion of SSR with respect to the individual length of each UTR, we observe large variations in the repeat content. It is possible that

some relationship exists between individual UTR length and SSR content because the demonstrations of post-transcriptional activities of UTR in trypanosomatids have been mapped mostly in 3′ UTR and also with involvement of SSR like sequences [4–6,9,21,22,29–33]. Even though having most of its genes intronless like *Giardia lamblia* and ranked at the lower positions in the Tree of Life [12], *T. cruzi* presents a distribution of 5′ UTR size similar to the intronic organisms. According to postulations on the origin of gene structure [34], organisms with primitive machineries for RNA processing or unable to perform RNA control like the NMD pathway should not possess 5′ UTR with such organization. Lynch argues in his model of 5′ UTR expansion that there is a limit posed onto the length of the 5′ UTR by the number of upstream ATG. But this model ignores a process demonstrated in trypanosomatids — the simultaneous occurrence of the 5′ capping (mini-exon addition) and 3′ polyadenylation of the pre-mRNA [35]. Though not exclusive to the trypanosomatids (in nematodes, for example, the trans-splicing has been shown to occur in at least 20% of genes [36]), the *quasi* intronless nature of trypanosomatids genome adds peculiarity to this phenomenon. In *T. cruzi*, the requirement of signaling sequences like the pyrimidine rich region and spliced leader addition site indicates that the 5′ UTR cannot be so short, even for those genes that encode small proteins such as the ribosomal proteins. Other mechanisms are certainly operating to maintain in *T. cruzi* a transcriptional machinery that resemble in shape the prokaryote one but are complex enough to depend on the structures of typical eukaryotes (large 5′ UTR, multiple transcription initiation sites, UTR control of gene expression and translation). In conclusion, the condition of a primitive eukaryote is to some extent mirrored on the *T. cruzi* UTRs, since some features typical of multicellular eukaryotes UTRs are not detected in its composition profile. However, in contrast to another primitive eukaryote *G. lamblia* [37], *T. cruzi* UTRs do not deviate too much from a typical eukaryote. It is thus a good snapshot model organism for the analysis of UTR function in the evolutionary transition from intronless to intronic genome.

Acknowledgements

This work is supported by the Academy of Sciences for the Developing World (formerly TWAS — http://www.twas.org) through a TWAS-Unesco Associateship scheme to Adeilton Brandão and the Bai Ren Project of the Chinese Academy of Sciences (http://www.cas.ac.cn), Project 973 (Grants: 2009CB918503 and 2006CB911002) to Taijiao Jiang.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.parint.2009.06.001.

References

[1] Simpson AGB, Stevens JR, Lukeš J. The evolution and diversity of kinetoplastid flagellates. Trends in Parasitology 2006;22:168–74.
[2] Jones SJM. Prediction of genomic functional elements annual review of genomics and human. Genetics 2006;7:315–38.
[3] Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene 2001;276:73–81.
[4] Effects of 3′ untranslated and intergenic regions on gene expression in *Trypanosoma cruzi*. In: Nozaki T, Cross GA, editors. Mol Biochem Parasitol 1995;75:55–67.
[5] Weston D, La Flamme AC, Van Voorhis WC. Expression of *Trypanosoma cruzi* surface antigen FL-160 is controlled by elements in the 3′ untranslated, the 3′ intergenic, and the coding regions. Mol Biochem Parasitol 1999;102:53–66.
[6] Coughlin BC, Teixeira SM, Kirchhoff LV, Donelson JE. Amastin mRNA abundance in *Trypanosoma cruzi* is controlled by a 3′-untranslated region position-dependent cis-element and an untranslated region-binding protein. J Biol Chem 2000;275:12051–60.
[7] Di Noia JM, D'Orso I, Sánchez DO, Frasch AC. AU-rich elements in the 3′-untranslated region of a new mucin-type gene family of *Trypanosoma cruzi* confers mRNA instability and modulates translation efficiency. J Biol Chem 2000;275:10218–27.
[8] Bartholomeu DC, Silva RA, Galvão LM, Elsayed NM, Donelson JE, Teixeira SM. *Trypanosoma cruzi*: RNA structure and post-transcriptional control of tubulin gene expression. Exp Parasitol 2002;102:123–33.
[9] Jäger AV, Muiá RP, Campetella O. Stage-specific expression of *Trypanosoma cruzi* trans-sialidase involves highly conserved 3′ untranslated regions. FEMS Microbiol Lett 2008;283:182–8.
[10] Brandão A. The untranslated regions of genes from *Trypanosoma cruzi*: perspectives for functional characterization of strains and isolates. Mem Inst Oswaldo Cruz 2006;101:775–7.
[11] Campos PC, Bartholomeu DC, DaRocha WD, Cerqueira GC, Teixeira SM. Sequences involved in mRNA processing in *Trypanosoma cruzi*. Int J Parasitol 2008;38:1383–9.
[12] Maddison DR, Schulz KS, (eds). 2007. The Tree of Life Web Project: http://tolweb.org/tree/.
[13] Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biology 2002;3:00041–000410 reviews.
[14] McClelland M, Ivarie R. Asymmetrical distribution of CpG in an 'average' mammalian gene. Nucleic Acids Res 1982;10:7865–77.
[15] Mrazek J, Gaynon LH, Karlin S. Frequent oligonucleotide motifs in genomes of three streptococci. Nucleic Acids Res 2002;30:4216–21.
[16] Kolpakov R, Bana G, Kucherov G. mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 2003;31:3672–8.
[17] Gautier C. Compositional bias in DNA. Curr Opin Genet Dev 2000;10:656–61.
[18] Hammer Ø, Harper DAT, Ryan PD. PAST: Paleontological Statistics Software Package for Education and Data Analysis Palaeontologia Electronica 2001; 4: 1, 9pp Available: http://palaeo-electronicaorg/2001_1/past/issue1_01htm Accessed December 03, 2008.
[19] Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis; MEGA software version 40. Mol Biol Evol 2007;24:1596–9.
[20] Vazquez M, Ben-Dov C, Lorenzi H, Moore T, Schijman A, Levin MJ. The short interspersed repetitive element of *Trypanosoma cruzi*, SIRE, is part of VIPER, an unusual retroelement related to long terminal repeat retrotransposons. Proc Natl Acad Sci USA 2000;97:2128–33.
[21] Boucher N, Wu Y, Dumas C, Dube M, Sereno D, Breton M, et al. A common mechanism of stage-regulated gene expression in *Leishmania* mediated by a conserved 3′-untranslated region element. J Biol Chem 2002;277:19511–20.
[22] McNicoll F, Müller M, Cloutier S, Boilard N, Rochette A, Dubé M, et al. Distinct 3′-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*. J Biol Chem 2005;280:35238–46.
[23] Mayho M, Fenn K, Craddy P, Crosthwaite S, Matthews K. Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in *Trypanosoma brucei*: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements. Nucleic Acids Res 2006;34:5312–24.
[24] Davuluri RV, Suzuki Y, Sugano S, Zhang MQ. CART classification of human 5′ UTR sequences. Genome Res 2000;10:1807–16.
[25] Brandão A. Trypanosomatid EST: a neglected information resource regarding flagellated protozoa? Mem Inst Oswaldo Cruz 2008;103:622–6.
[26] Fernandes O, Santos SS, Cupolillo E, Mendonça B, Derre R, Junqueira AC, et al. A mini-exon multiplex polymerase chain reaction to distinguish the major groups of *Trypanosoma cruzi* and *T. rangeli* in the Brazilian Amazon. Trans R Soc Trop Med Hyg 2001;95:97–9.
[27] Brandão A, Fernandes O. *Trypanosoma cruzi*: mutations in the 3′ untranslated region of calmodulin gene are specific for lineages *T. cruzi* I, *T. cruzi* II, and the Zymodeme III isolates. Exp Parasitol 2006;112:247–52.
[28] Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 2004;2:991–1007.
[29] Furger A, Schürch N, Kurath U, Roditi I. Elements in the 3′ untranslated region of procyclin mRNA regulate expression in insect forms of *Trypanosoma brucei* by modulating RNA stability and translation. Mol Cell Biol 1997;17:4372–80.
[30] Lee MG. The 3′ untranslated region of the hsp 70 genes maintains the level of steady state mRNA in *Trypanosoma brucei* upon heat shock. Nucleic Acids Res 1998;26:4025–33.
[31] Quijada L, Soto M, Alonso C, Requena JM. Identification of a putative regulatory element in the 3′-untranslated region that controls expression of HSP70 in *Leishmania infantum*. Mol Biochem Parasitol 2000;110:79–91.
[32] Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M. Developmental regulation of heat shock protein 83 in *Leishmania*: 3′ processing and mRNA stability control transcript abundance, and translation id directed by a determinant in the 3′-untranslated region. J Biol Chem 2001;276:47922–9.
[33] Irmer H, Clayton C. Degradation of the unstable EP1 mRNA in *Trypanosoma brucei* involves initial destruction of the 3′-untranslated region. Nucleic Acids Res 2001;29:4707–15.
[34] Lynch M. The origins of eukaryotic gene structure. Mol Biol Evol 2006;23:450–68.
[35] LeBowitz JH, Smith HQ, Rusche L, Beverley SM. Coupling of poly(A) site selection and trans-splicing in *Leishmania*. Genes Dev 1993;7:996–1007.
[36] Guiliano DB, Blaxter ML. Operon conservation and the evolution of trans-splicing in the phylum Nematoda. PLoS Genet 2006;2:e198.
[37] Li L, Wang CC. Capped mRNA with a single nucleotide leader is optimally translated in a primitive eukaryote, *Giardia lamblia*. J Biol Chem 2004;279:14656–64.