Anais da Academia Brasileira de Ciências (2018) 90(2 Suppl. 1): 2311-2316
(Annals of the Brazilian Academy of Sciences)
Printed version ISSN 0001-3765 / Online version ISSN 1678-2690
http://dx.doi.org/10.1590/0001-3765201720170082
www.scielo.br/aabc | www.fb.com/aabcjournal

Minimum free energy predicted base pairing in the 39 nt spliced leader and 5' UTR of calmodulin mRNA from *Trypanosoma cruzi*: influence of the multiple trans-splicing sites

FRANKLYN SAMUDIO[1,2] and ADEILTON BRANDÃO[1]

[1]Laboratório Interdisciplinar de Pesquisas Médicas, Instituto Oswaldo Cruz, Fiocruz, Av. Brasil, 4365, 21040-360 Rio de Janeiro, RJ, Brazil
[2]Instituto Conmemorativo Gorgas de Estudios de la Salud, Ave. Justo Arosemena, Entre calle 35 y 36, Apartado Postal 0816-02593, Panamá, República de Panamá

Manuscript received on February 7, 2017; accepted for publication on June 29, 2017

ABSTRACT

We analyzed the compositional changes and the stable base pairs in the predicted secondary structure of the 5' UTR calmodulin mRNA in *T. cruzi*. The three copies of calmodulin in *T. cruzi* genome display variable position of the trans splicing sites and give rise to several mRNA that differs slightly on 5' UTR composition in the epimastigote stage. We show that the pattern of high probability base pairs in the minimum free energy predicted secondary structures of the calmodulin 5' UTR remains unchanged despite the nucleotide composition variation. However, the 39 nt spliced leader (mini-exon, the 5' exon sequence transferred to trypanosome mRNAs by the mechanism of trans splicing) shows a variable pattern of high and low probability base pairing as consequence of the altered composition of the 5' UTR.

Key words: trans-splicing, untranslated region, calmodulin, epimastigote, RNA secondary structure, minimum free energy.

INTRODUCTION

Though *T. cruzi* is typically a eukaryote, it displays unusual features like polycistronic transcription, absence of introns in most of the genes, and processing of protein-encoding RNA by the mechanism of trans-splicing coupled to polyadenylation (Araújo and Teixeira 2011). As parasite, *T. cruzi* accomplishes its life cycle by moving from invertebrate to vertebrate hosts (Coura 2014). This transition submits the parasite to physical and chemical shifts that include wide variations in temperature, nutrient availability, osmotic and environmental pressure. Trypanosome transcriptome analysis shows that the processing of the primary polycistronic transcript from tandem arrayed genes might generate several variants of the corresponding monocistronic mRNA, mainly due to the use of additional sites of trans-splicing and polyadenylation (Nilsson et al. 2010). These facts call attention to the problem of the alterations in the 5' UTR secondary structure in certain genes that might be caused by changes in length and composition of 5' UTR derived from distinct

Correspondence to: Adeilton Brandão
E-mail: abrandao@fiocruz.br

* Contribution to the centenary of the Brazilian Academy of Sciences.

trans splicing sites. To shed light on this problem, we analyzed in this work the effect on both the 5' UTR and the 39 nt SL (mini-exon, the 5' exon sequence transferred to trypanosome mRNAs by the mechanism of trans splicing) of the multiple trans splicing sites of the calmodulin gene in the epimastigote stage of *T. cruzi* Y strain.

MATERIALS AND METHODS

T. cruzi Y strain epimastigotes were cultivated at 28 ºC in brain hearth infusion (BHI) (Becton, Dickinson and Company) supplemented with of 10% of heat-inactivated fetal bovine serum (Life technologies). At least 2x 10^8 mid-log phase epimastigotes were recovered by centrifugation at 3,000 rpm for 10 min and washed three times with PBS. The RNA was extracted by Trizol (Life technologies) according to manufacturer instructions, and resuspended in 30 μL of water. Calmodulin 5' UTR was obtained by RT-PCR using primers targeted to the spliced leader (forward primer based on the first 25 bases of the 39 nt spliced leader: 5' AACTAACGCTATTATTGATACAGTT 3') and calmodulin 3' UTR common to all copies (reverse primer: 5' GTCACTGTCGGCTTCGCT 3'). Two micrograms of RNA were used to produce cDNA in a standard reaction at 50 °C for 90 min with 2 U of Super Script III (Life technologies) and 500 nM of the anchored primers T(13)A, T(13)G, T(13)C. Then a PCR was carried out using Platinum® Taq DNA Polymerase High Fidelity (Life technologies) and 100 nM of primers forward and reverse at 94°C for 2 min followed by 35 cycles of 97 °C for 15 s, 60 °C for 20 s, 72 °C for 30 s and a final extension of 72 °C for 7 min. Pyrosequencing were carried out at the high-throughput sequencing facility of the Instituto Oswaldo Cruz – Fiocruz. The sff file (Roche 454 GS Junior output) was converted to a fastQ file by the sff2fastq and further edited by program CANGS (Pandey et al. 2010). The sequences have been submitted to Sequence Read Archive (SRA) of Genbank-NCBI under accession codes: Experiment: SRX734426, Run: SRR1614234. Two hundred bases upstream of the start codon from each copy of the calmodulin locus from CL Brener Esmeraldo haplotype were used to generate sequence alignments through Clustal Omega in the UGENE package. Though the CL Brener clone and Y strain have been allocated to different *T. cruzi* DTUs (DTU VI for CL Brener and DUT II for Y strain) (Zingales et al. 2009), the tandem arrangement of the calmodulin gene is invariable in *T. cruzi* strains belonging to different DTU (Brandão and Fernandes 2006). Complete 5' UTR sequences (including the 39 nt SL) were submitted to the RNAfold in the viennaRNA Web Services under default parameters for the prediction of the base pair probabilities in the minimum free energy structure (Gruber et al. 2008).

RESULTS AND DISCUSSION

The calmodulin locus is organized as three tandem copies with 450 nt ORF (copy accession numbers: TcCLB.507483.50, TcCLB.507483.39, and TcCLB.507483.30). Five point mutations occur among the three ORF, but only one results in amino acid change. The 5' UTR in TcCLB.507483.50 is distinct from the two other copies but it is almost identical in copies TcCLB.507483.39 and TcCLB.507483.30. We have used the point mutation in nucleotide position 69 of the ORF to distinguish the origin of each mRNA in the last two copies. Figure 1 shows a diagram of the calmodulin locus in the *T. cruzi* CL Brener haplotype Esmeraldo. Two spacers of 1,388 and 702 nt separate the pairs TcCLB.507483.50 – TcCLB.507483.39 and TcCLB.507483.39 – TcCLB.507483.30, respectively.

The observed trans splicing sites in each copy were as follows (in bold):

a) TcCLB.507483.50: **4**;
b) TcCLB.507483.39: **8**;

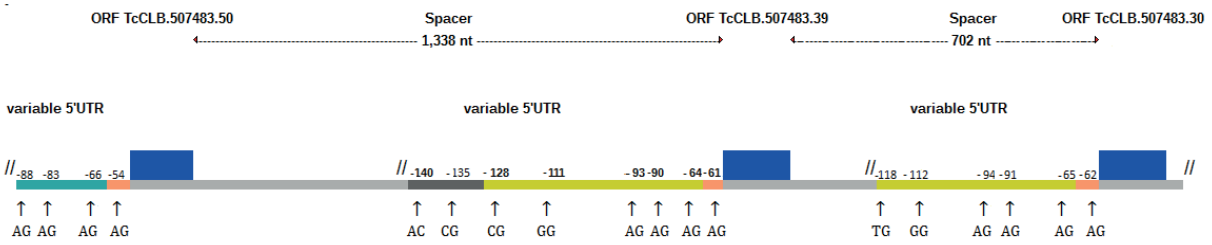

Figure 1 - Diagram of *T. cruzi* calmodulin locus and frequency of the mRNA variant sequences. The locus displays three tandem copies that are separated by two intergenic segments of 1,338 and 702 nt. Blue box: ORF; colored bars: 5' UTR that results from the multiple trans splicing sites as detected in epimastigote stage; in orange and yellow are the 5' UTR segment common to three or two copies, respectively. Arrows and numbers at left of the blue box indicate the approximate position of the trans splicing site (and the respective dinucleotide below the arrows), which also define the beginning of 5' UTR in the genome for each mRNA.

c) TcCLB.507483.30: **6**.

As expected, most of these sites are defined by the canonical dinucleotide AG, but non canonical trans splicing sites such as CA, CC, CG, GG and TG were observed as well.

The processing of polycistronic transcripts in trypanosomatids might result in abundance variation of the final mRNA of a particular gene or group of genes (Clayton 2014). According to transcriptome data provided by Li et al. (2016), the calmodulin expression level in epimastigote stage varies among its copies, indicated here are by average log-transformed quantile-normalized cpm expression values (in bold):

a) TcCLB.507483.50: **7,79**;
b) TcCLB.507483.39: **8,17**;
c) TcCLB.507483.30: **8,67**

These expression values are not linearly related to the number of trans splicing sites for each copy, although the copy with the lowest expression level (TcCLB.507483.50) is also the one with fewer trans splicing sites. This copy is located in one of the extremities of the calmodulin locus (see fig. 1).

A given RNA molecule can be viewed as a thermodynamic ensemble of structures and does not necessarily exhibit a unique secondary structure under cell physiological conditions (Ding and Lawrence 2003). Thus, instead of looking at the predicted secondary structure, it is more informative to look at the paired base segments that occur more frequently in the thermodynamical ensemble of low free energy structures. These high probability paired base segments are certainly to be present in the actual secondary structure for that RNA (Mathews et al. 2010). The multiple trans splicing sites shown to be present in calmodulin locus give rise to a wide range of length variation in the final mRNA, implicating in millions of alternative secondary structures for the 5' UTR. We asked whether the variable 5' UTR length is directly related to the degree of structured sequence in each mRNA from the three gene copies. To address this question, we used a reliability measure of the RNAfold software: colorized base pairing probabilities for each 5' UTR. As shown in fig. 3, the 39 nt spliced leader (shown partially in black line rectangle) may assume different configurations with a variable number of base pairs. The most stable ones (high probability segments in red) occur in the 5' UTR of intermediate length (104 – 150 nt). The paired bases (stems) of the spliced leader are closer to beginning of the 5' UTR. However, most of the paired bases are of low probability. It is noteworthy how the predicted mfe structure of 39 nt SL changes with the composition of the 5' UTR, which is a consequence of the "movable" trans splicing site. As shown in figure 2, the mfe predicted secondary structure of the SL sequence alone (39 nt) displays

Figure 2 - Predicted mfe secondary structure with base pairing probabilities for the 39 nt spliced leader present in mRNA from *T. cruzi*. The sequence was folded by RNAfold under default parameters.

a single stem loop with 5 bp helix. In this predicted structure, the highest probability is assigned to the unpaired bases at the 5' end (the 12 single bases in red). This structure is highly improbable *in vivo* because the 39 nt SL does not exist isolated in the cell: either it is coupled to original donor RNA or to the 5' end of messenger RNA. By comparing this same 39 nt SL segment coupled to the variable 5' UTR segments, another picture emerges from the predicted structures: the single stem loop disappears to give rise to a variable number of low and high probability short helices (paired bases). In all 5' UTR analyzed here, only the first 5 - 10 nucleotides at 5' end of SL appears unpaired at high probabilities. Most of the remaining bases form either low or high probability base pairing with 5' segments of the various 5' UTR. In this specific case, the 39 nt spliced leader adapts its secondary structure to the composition of the 5' UTR. This implies an importance to composition alterations in the 5' UTR from *T. cruzi*, in particular those caused by changes in the trans splicing site.

In the mid 1990s, researchers have shown that SL from *Leptomonas collosoma* and *T. brucei* displayed two main structures (LeCuyer and Crothers 1993). Similar work *in vivo* (permeabilized cells) demonstrated that the SL alternates between these structures with predominance of structure 1 (Harris et al. 1995). Although some of the predicted SL structures in the 5' UTR analyzed here roughly approached the structures characterized earlier, these results should be interpreted with caution because the computer prediction of RNA structures does not exclude the possibility that less thermodynamically stable structure are the real biological (*in vivo*) structures. What the prediction allows us to ascertain is that only the first 5-8 bases of SL remain single base at high probability for most of the 5' UTR sequences in calmodulin mRNA. Considering that millions of structures might exist beyond the most thermodynamically stable (mfe structure), it cannot be ruled out that structures not selected by the current algorithms are completely absent of any functional activity *in vivo*.

According to mfe predictions presented here, the thermodynamically stable segments in the 5' UTR of calmodulin mRNA do not change the amount of paired bases with the additional trans splicing sites used to process the primary mRNA. It is also worth noting that the different length and composition of the 5' UTR do not affect the formation of a stem loop in the high probability 41 nt segment close to the calmodulin start codon (black line circle in fig. 3).

With the exception of the stem loop close to start codon, the pairing probability for *T. cruzi* calmodulin 5' UTR sequences in the epimastigote stage shows that most of these segments remain single base, *i. e.*, do not form a secondary structure. For the 5' UTR this is not a surprise, as the 5' leader sequence in the mRNA is expected to be lightly structured, unless a large portion of the 5' UTR is involved in gene expression control and modulation of translation rates (Hinnebusch et al. 2016).

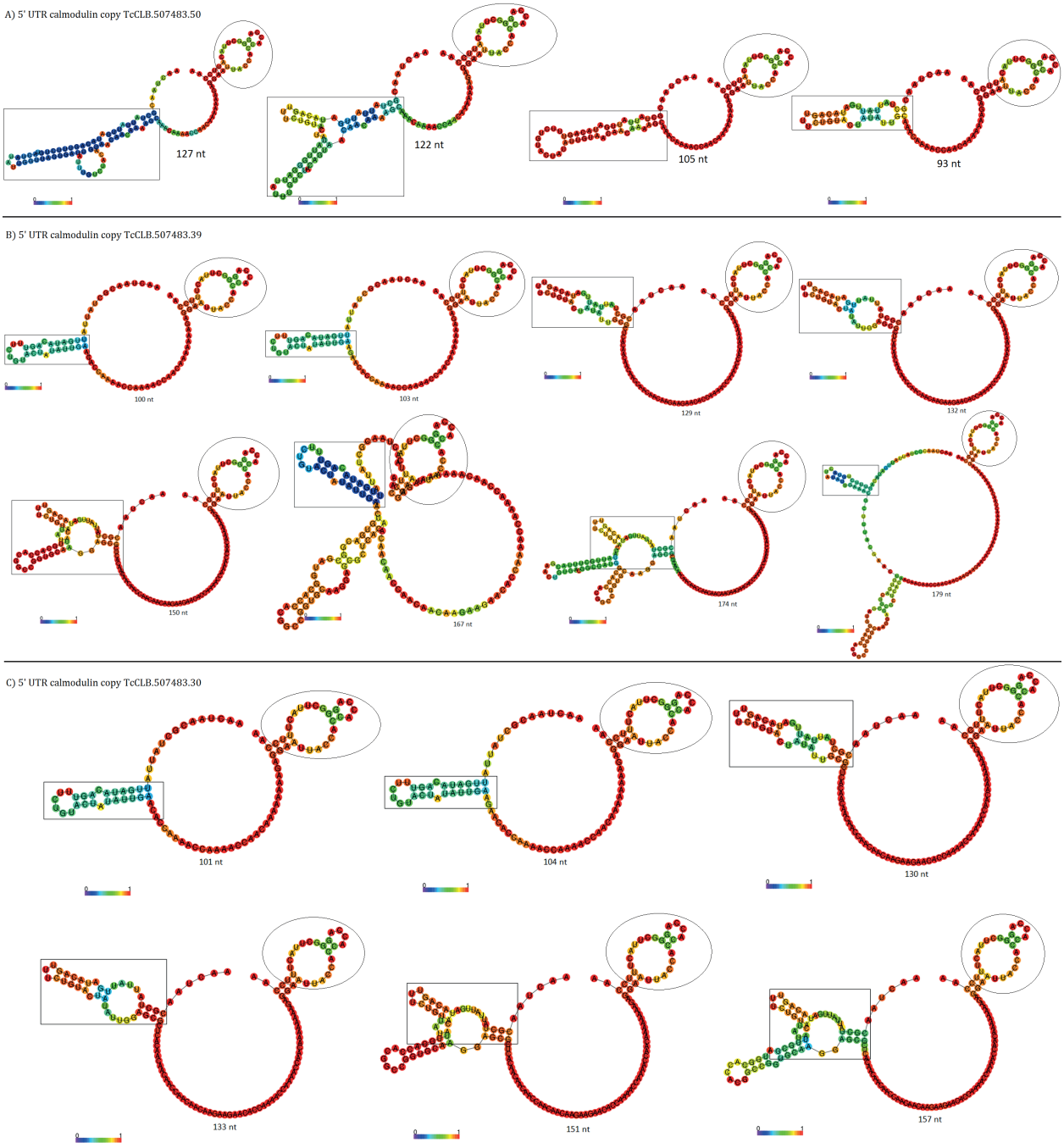

Figure 3 - Predicted mfe structure with base pairing probabilities for the calmodulin 5' UTR in *T. cruzi* epimastigote. The 5' UTR sequences (grouped by copy origin) were joined to 39 nt SL (black line rectangle showing partially paired segment) and folded by RNAfold. Colors in the schematic structure refer to probability of base pairing or unpairing (remaining as single base). Red segments contain the paired (or unpaired) bases of the highest probability. Inside the black line rectangle are the paired bases that the 39 nt spliced leader forms with either itself or the 5' UTR sequence. Most of these SL paired bases are of low probability. Black line circles indicate the stem loop with high probability paired bases near the start codon of the calmodulin mRNA. Numbers below the major single base segment denote the 5' UTR length (plus the spliced leader).

CONCLUSIONS

Despite alterations in nucleotide composition and mRNA expression level for the three gene copies, the calmodulin 5' UTR tends to maintain roughly the same pattern of stable paired/ unpaired bases. In contrast, the 39 nt spliced leader sequence shows a variable segment of high and low probability paired bases.

ACKNOWLEDGMENTS

This work received financial support from Gorgas Institute and from SENACYT, Panama (grant COL08-080), the IFARHU-SENACYT, Panama, and PROEP/CNPq/IOC/Fiocruz, Brazil.

REFERENCES

ARAÚJO PR AND TEIXEIRA SM. 2011. Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in *Trypanosoma cruzi*: a review. Mem Inst Oswaldo Cruz 106: 257-266.

BRANDÃO A AND FERNANDES O. 2006. *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 112: 247-252.

CLAYTON CE. 2014. Networks of Gene Expression Regulation in *Trypanosoma brucei*. Mol Biochem Parasitol 195: 96-106.

COURA JR. 2014. The main scenarios of Chagas disease transmission. The vectors, blood and oral transmissions - A comprehensive review. Mem Inst Oswaldo Cruz 110: 277-282.

DING Y AND LAWRENCE CE. 2003. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 31: 7280-7301.

GRUBER AR, LORENZ R, BERNHART SH, NEUBÖCK R AND HOFACKER IL. 2008. The Vienna RNA Web suite. Nucleic Acids Res 36(suppl 2): W70-W74.

HARRIS KA JR, CROTHERS DM AND ULLU E. 1995. In vivo structural analysis of spliced leader RNAs in *Trypanosoma brucei* and *Leptomonas collosoma*: a flexible structure that is independent of cap4 methylations. RNA 1: 351-362.

HINNEBUSCH AG, IVANOV IP AND SONENBERG N. 2016. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science 352: 1413-1416.

LECUYER KA AND CROTHERS DM. 1993. The *Leptomonas collosoma* spliced leader RNA can switch between two alternate structural forms. Biochemistry 32: 5301-5311.

LI Y ET AL. 2016. Transcriptome Remodeling in *Trypanosoma cruzi* and Human Cells during Intracellular Infection. PLOS Pathogens 12(4): e1005511.

MATHEWS DH, MOSS WN AND TURNER DH. 2010. Folding and Finding RNA Secondary Structure. Cold Spring Harbor Perspectives in Biology 2: a003665.

NILSSON D, GUNASEKERA K, MANI J, OSTERAS M, FARINELLI L, BAERLOCHER L, RODITI I AND OCHSENREITER T. 2010. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. PLoS Pathogens 6: e1001037.

PANDEY RV, NOLTE V AND SCHLÖTTERER C. 2010. CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. BMC Research Notes 3: 3.

ZINGALES B ET AL. 2009. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Mem Inst Oswaldo Cruz 104: 1051-1054.