Available online at www.sciencedirect.com
SCIENCE DIRECT®
Experimental Parasitology 109 (2005) 143–149
Experimental Parasitology
www.elsevier.com/locate/yexpr

Strand asymmetry patterns in trypanosomatid parasites

Daniel Nilsson, Björn Andersson*

Center for Genomics and Bioinformatics, Karolinska Institutet, Berzeliusv. 35, SE-171 77 Stockholm, Sweden

Received 12 August 2002; received in revised form 1 December 2004; accepted 1 December 2004

Abstract

The genome organization of kinetoplastid parasites is unusual, with chromosomes containing several long regions of polycistronically transcribed genes. The regions where the direction of transcription switches have been hypothesized to contain origins of replication and possibly also centromers and promoters. We report that overall strand asymmetry patterns can be observed in *Trypanosoma cruzi* and *Trypanosoma brucei* with optima on strand-switch regions. The base skews of *T. cruzi* and *T. brucei* divergent strand-switches show patterns analogous to those for bacterial origins of replication, but they differ from those of *Leishmania major*. Bias in codon usage and the trypanosomatid unidirectional gene clusters predict most of this skew, but fail to properly explain the same trend in intergenic regions, as does the current knowledge of regulatory sequences.

Index Descriptors and Abbreviations: Parasite genomics; Parity rule two deviation; Strand asymmetry; Codon usage; Trypanosomatids; *Lm*, *Leishmania major* Friedlin; *Tb*, *Trypanosoma brucei*; *Tc*, *Trypanosoma cruzi*; Ori, Origin of replication; Ter, Termini of replication

1. Introduction

Studies of trypanosomatid parasites have revealed surprising phenomena in the past, such as *trans*-splicing and RNA editing. The first longer contiguous genomic sequences from the genome projects in *Leishmania major* Friedlin (*Lm*), *Trypanosoma brucei* (*Tb*), and *Trypanosoma cruzi* (*Tc*) (reviewed in Degrave et al., 2001) have begun to reveal the genome organization of these parasites. Trypanosomatid genes are almost exclusively encoded in unidirectional clusters, with intervening strand-switches (El-Sayed et al., 2003; Hall et al., 2003; Worthey et al., 2003). The genes in a cluster are transcribed polycistronically and *trans*-spliced to form mature mRNA molecules from each gene. The strand-switches have attracted considerable interest, and have been suggested to contain regulatory elements, transcriptional promoters (Wong et al., 1994), origins of replication, as well as to constitute centromers (Myler et al., 1999) of the trypanosomatid chromosomes, but no conclusive evidence for any of these elements has as yet been presented.

Analyses of genomic sequences have revealed distinct patterns of deviation from Chargaff's second rule of parity (PR2) (Lobry, 1996; Wu and Maeda, 1987). PR2 states that the intra-strand nucleotide concentration [C] = [G], as well as [A] = [T], under equilibrium assumptions given Watson–Crick basepairing and no bias in selection or mutation between the strands (Lobry, 1995; Sueoka, 1995). Many eubacterial genomes show a strong correlation between coding excess (gene-orientation) and the degree of deviation from PR2, quantified as GC-skew and AT-skew, with optima in the cumulative measures precisely at the origin (ori) and terminus (ter) of replication (Grigoriev, 1998). The GC-skew has proven to be more predictive than the AT-skew in this regard. Also, plots of purine excess have been demonstrated to correlate with ori and ter in some, but not all, cases (Freeman and Plasterer, 1998). The situation in eukaryotes is less clear, with many local skew

* Corresponding author. Fax: +46 8 311620.
E-mail address: bjorn.andersson@cgb.ki.se (B. Andersson).
URL: http://cruzi.cgb.ki.se/ (B. Andersson).

0014-4894/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.exppara.2004.12.004

optima (minima and maxima), that have been postulated to coincide with multiple ori and ter (Shioiri and Takahata, 2001). Recent findings support this (Niu et al., 2003).

Many alternative explanations for the origin of these patterns have been proposed (see, e.g., Frank and Lobry, 1999, for a review). Several cellular processes are asymmetric with respect to DNA strands and thus constitute potential causes of asymmetric base distributions.

Errors introduced in transcription coupled repair of the coding strand have been identified as causative of base skew in certain organisms (Francino and Ochman, 1997; Green et al., 2003). In transcription, the coding strand (non-template) is more unprotected, whereas the non-coding (template) strand is not only more protected, but also undergoes correcting excision repair to a greater extent (Mellon and Hanawalt, 1989). Also, during replication, the strands are copied by different systems of enzymes, and face different risks of DNA damage, as well as different repair mechanisms (Fijalkowska et al., 1998; Izuta et al., 1995; Maliszewska-Tkaczyk et al., 2000). This can lead to different fidelity of replication for the leading and lagging strands (Reyes et al., 1998; Rocha and Danchin, 2001). At least in eubacteria, the direction of replication and transcription is often correlated (Brewer, 1988; McLean et al., 1998) so that a skew introduced by either process can be added to that of the other.

Regular patterns of purine excess and GC- and AT-skew have been found in *Lm* chromosome 1 (McDonagh et al., 2000), surprisingly showing a negative correlation between coding excess and GC-skew, which is the opposite of the pattern found in eubacteria. The analysis of *Lm* led McDonagh et al. to suggest that the effects of transcription coupled repair are balanced by ubiquitous and strand un-specific transcription, which has been found to occur in kinetoplastids (Clayton, 2002).

This interpretation now seems unlikely given results from studies on tentative promoter activity in trypanosomes that suggest more strand-specific transcription (Martinez-Calvillo et al., 2003), from the non-coding (template) strand only, as in many other organisms. This view is complicated by results that show a background level of transcription initiation independent of promoters (Clayton, 2002, and references therein). It is still unclear how extensive this background transcription is, and if all regions of the genome behave equally in this regard.

We find that the skew resulting from unequal codon usage explains much of the correlation between coding excess and skew seen in the three parasites investigated, but that it is clearly present also in the third codon position, and furthermore, that this pattern is present in intergenic regions as well. Intriguingly, the chromosomes of *Tb* and *Tc* show a positive correlation between coding excess and GC-skew—the opposite of that of the closely related *Lm*.

2. Methods

The sequences used were *Tb* chromosome I (GenBank Accession No. AL359782), *Tc* chromosome 3 (AF052831, AF052832, and AF052833), and *Lm* chromosome 1 (NC_001905).

The sequence was subdivided into coding and non-coding according to the coordinates given in GenBank CDS entries. All calculations were performed on the 5′–3′ strand exclusively, compensating for the codon position of any "reverse"-strand coding regions accordingly. The strand-switches were set to be at nucleotide position 24,500 for *Tc*, 78,000 for *Lm*, and 243,000 for *Tb*.

The cumulative GC (or AT) base skew s_{RY} was calculated as

$$s^{j}_{\mathrm{RY}} = \sum_{n=0}^{j} \sum_{i=1}^{l} \frac{\delta^{n+i}_{\mathrm{R}} - \delta^{n+i}_{\mathrm{Y}}}{\delta^{n+i}_{\mathrm{R}} + \delta^{n+i}_{\mathrm{Y}}},$$

where N is the sequence length, l is the window length, and $j \leqslant N - l$ is the sequence position. Arbitrary overlapping window lengths of 400 and 10,000 bp have been used. δ^{x}_{R} is 1 for position x only if there is a purine R at that position and zero otherwise; δ^{x}_{Y} analogously for pyrimidine Y.

Coding excess was calculated as

$$s_{\mathrm{ce}} = \sum_{n=1}^{N} \delta^{n}_{\mathrm{F}} - \delta^{n}_{\mathrm{R}},$$

where δ^{n}_{F} is one if and only if base n is in a gene encoded on the forward strand, and 0 otherwise. δ^{n}_{R}, analogously, takes the value 1 only for bases in genes encoded on the reverse strand, and 0 otherwise. Thus s_{ce} is a cumulative measure of which strand the genes are encoded on, with optima at strand-switches.

The skew contributions for each of the 64 codons were calculated, and weighted with the codon usage (as estimated by Nakamura et al., 1999) for each of the three parasites. This is the expected average skew per nucleotide given uniform amino-acid and stop codon usage. We note that the codon usage estimations rely in part on data from the sequences subsequently analysed in this work. The sequences analysed contribute only a small fraction of the CDS entries used in the codon usage estimation for *Tc* (39/420) and *Lm* (79/1426), but for *Tb* the fraction is relatively large (325/422).

The tandem repeat analyses were executed as in Duhagon et al. (2001), locating exact repeats using a perl (Wall and Schwartz, 1991) program and with the