# Pre-mRNA *trans*-splicing: from kinetoplastids to mammals, an easy language for life diversity

Mario Gustavo Mayer , Lucile Maria Floeter-Winter[*/+]

Laboratório de Genética, Instituto Butantan, São Paulo, SP, Brasil *Departamento de Fisiologia, Instituto de Biociências, Rua do Matão, travessa 14, 101, 05508-900 São Paulo, SP, Brasil

*Since the discovery that genes are split into intron and exons, the studies of the mechanisms involved in splicing pointed to presence of consensus signals in an attempt to generalize the process for all living cells. However, as discussed in the present review, splicing is a theme full of variations. The* trans-*splicing of pre-mRNAs, the joining of exons from distinct transcripts, is one of these variations with broad distribution in the phylogenetic tree. The biological meaning of this phenomenon is discussed encompassing reactions resembling a possible noise to mechanisms of gene expression regulation. All of them however, can contribute to the generation of life diversity.*

Key words: RNA processing - alternative *trans*-splicing - spliced leader RNA - trypanosomatids - nematodes - mammalian

Most of the protein-coding eukaryotic genes display an interrupted structure alternating exons and introns. After transcription, introns must be removed from the primary transcript (pre-mRNA) to generate a translatable mature mRNA. It is interesting to note that the mature mRNA is constituted of 5' and 3' untranslatable regions (UTR) flanking the open reading frame (ORF) and that UTR are also exons. The precise excision of the introns and the joining of neighboring exons is a complex process generally named splicing, and when this processing occurs within a single pre-mRNA molecule it can also be called *cis*-splicing (Moore et al. 1993, Burge et al. 1999).

*Cis*-splicing occurs in a two-step mechanism, each step consisting of a transesterification reaction (Moore et al. 1993) (Fig. 1A). The spliceosome, a ribonucleoprotein machinery composed of five small ribonucleoproteins (U1, U2, U4, U5, U6 snRNPs, named in relation to the kind of associated RNA molecule) and approximately 300 distinct proteins, is responsible for the splicing catalysis (Burge et al. 1999).

Much effort was expended on the comprehension of the structure and dynamics of this complex machine during the splicing process, but the rules that discriminate introns and exons within the message are still poorly understood. Nevertheless, four conserved pre-mRNA sequence elements, which interact with the spliceosome, have been characterized as determinants in the splicing process (Moore et al. 1993, Burge et al. 1999) (Fig. 1C). In mammals, the 5' splice site consensus is AG/**GU**RAGU (R for purine, /for splice site, bold for the dinucleotide 5' intron boundary) while the 3' splice site consensus is Y**AG**/G (Y for pyrimidine, /for splice site, bold for the dinucleotides 3' intron boundary). There is also a conservation of nucleotide sequence around the branch point, CUR**A**Y (Y for pyrimidine, R for purine and **A** for branch

point), and a polypyrimidine tract of variable length between the branch point and the 3' splice site.

Recently, other *cis*-regulatory elements were implicated in splice site recognition, e.g. the exonic splicing enhancers (ESEs) (Fig. 1C). ESEs are localized within exons and interact with a family of proteins rich in serine and arginine (SR proteins) that recruits the spliceosome to the proximal splice site (Maniatis & Tasic 2002).

Although the presence of those consensuses was observed, it is well known now that the splicing theme is full of variations. Yeast pre-mRNAs have a slightly different consensus for the 5' splice site (AG/**GU**AGU) than mammals, and the polypyrimidine tract is not as evident. Moreover, the discovery of other variations has broken some rules. Some introns, known as AU-AC, bear AU and AC at their 5' and 3' boundaries instead of the GU and AG dinucleotides, as well as different sequences around their branch point.

Variations are not restricted to pre-mRNA sequences and involve the spliceosome machinery itself. A minor spliceosome presents in its composition U11, U12, U4$_{atac}$ and U6$_{atac}$ snRNPs instead of U1, U2, U4, and U6 snRNPs always found in major spliceosomes. The only snRNP that is common to both types of spliceosome is the U5 snRNP (Yu et al. 1999).

The classical splicing mechanism involves the joining of exons encompassed in one pre-mRNA molecule (*cis*-splicing). One important variation of this picture can be found in early branched and at least in some pre-mRNAs of complex eukaryotes, where exons from distinct molecules could be joined together (*trans*-splicing) (Ullu et al. 1996). The following sections of the present communication intend to summarize the history, distribution, and possible biological meaning of the pre-mRNA *trans*-splicing process amongst different organisms.

Trans-*splicing: a brief history* **-** In vitro studies using HeLa cell-free system showed for the first time that two independent transcripts could be joined together by *trans*-splicing (Konarska et al. 1985, Solnick 1985). Those data promptly suggested that the same *trans* reaction could take place in vivo in eukaryotic cells.

The studies of antigenic variation in infective forms of the early-branched eukaryote *Trypanosoma brucei*, the causative agent of African trypanosomiasis, showed that all mRNAs encoding variant surface glycoproteins (VSGs) had an identical 39 nucleotide leader sequence at its 5' termini. That sequence, however, was absent in VSG genomic vicinities (Boothroyd & Cross 1982). Subsequent experiments identified the same sequence at the 5' termini in other mRNAs from *T. brucei* showing that the occurrence was not restricted to VSG genes, but should represent the addition of a leader sequence in all mRNA of that organism (De Lange et al. 1984). Further, similar sequences were found at 5' end of other trypanosomatids, showing that this phenomenon occurs in all those organisms (Parsons et al. 1984).

Further research showed that the leader sequence is present in approximately 200 copies per genome, clustered *in tandem*, which do not map in the same chromosome as VSG genes (De Lange et al. 1983). Expression studies showed that transcription of the leader sequences generates a small RNA of 135-147 nt, which contains the 39 nt leader sequence at its 5' termini (Milhausen et al. 1984). Although these results exclude the occurrence of *cis*-splicing, other hypotheses were proposed to explain the structure of mature mRNAs of trypanosomes: (i) the pre-mRNA is transcribed using the leader sequence as initiator, and the resulting molecule is then processed by *cis*-splicing; (ii) the independently transcribed short RNA bearing the leader sequence is ligated, via its 3' end to the 5' end of the pre-mRNA and the resulting molecule is

spliced in *cis*; (iii) the leader RNA and the pre-mRNA could be independently transcribed as in (ii) but the two substrates could be spliced in *trans*. Experimental analysis of pre-mRNA maturation detected a Y-branched intermediate molecule, instead of the typical lariat expected from *cis*-splicing, only compatible with a *trans*-splicing process (Fig. 1B) (Murphy et al. 1986, Sutton & Boothroyd 1986). The leader sequence was then named spliced leader (SL) or mini-exon, while the short transcript in the unprocessed form was named SL-RNA.

The *trans*-splicing processing in trypanosomes is very similar to the canonical *cis*-splicing (Fig. 1A, B) (Ullu et al. 1996). The SL RNA is one of the substrates and presents the canonical GU splice junction at the 39 nt exon intron boundary. The other substrate is the pre-mRNA that contains the 3' canonical AG splice site localized in the nontranslated 5' part of the molecule (Fig. 1B). This 5' region of the pre-mRNA also contains a conserved adenosine (A) residue that is used as the branch point in the process. Finally, a polypyrimidine tract is observed between the 3' acceptor site and the branch point.

The discovery of natural *trans*-splicing in trypanosomes raised the question of how common this kind of processing could be among other organisms. In fact, just after those studies, *trans*-splicing by spliced leader addition was detected, for at least part of nematodes transcripts (Krause & Hirsh 1987). Nevertheless, it was not known if this kind of processing was part of the pre-mRNA processing repertoire of other eukaryotes. Computational search of the known features of *trans*-splicing substrates
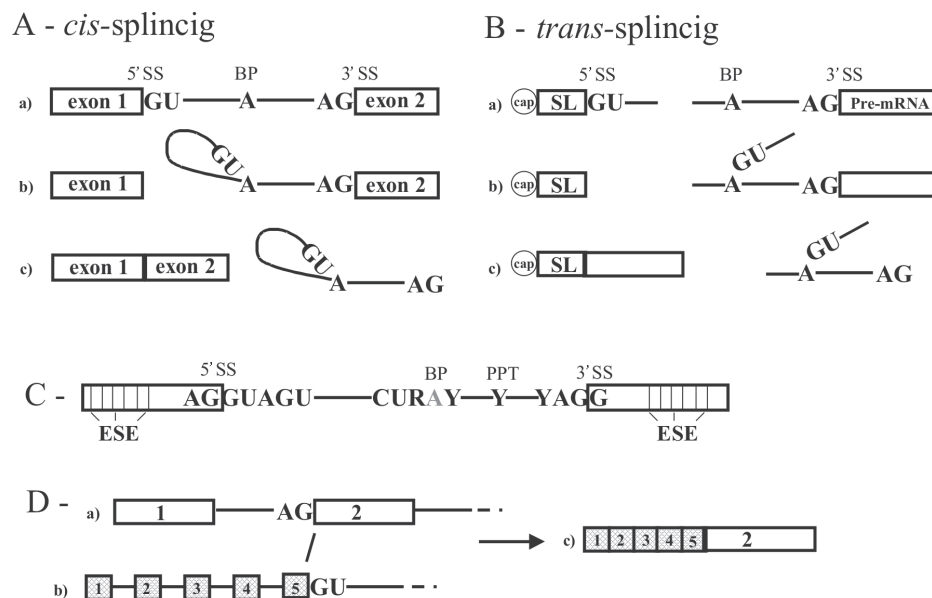


Fig. 1: schematic representation of *cis* and *trans*-splicing pathways. A: *cis*-splicing occurs through a two-step mechanism, each step consisting of a transesterification reaction (Moore et al. 1993). In the first transesterification (a to b), a nucleophilic attack of a 2'-OH group from an adenosine residue within the intron (branch point, BP) to the phosphorous atom at the 5' splice site (5' SS) generates two intermediates: exon 1 with a free 3'-OH terminus and an intron-exon 2 in a lariat configuration. In the second transesterification (b to c), the 3'-OH group form the exon 1 intermediate acts as a nucleophile and attacks the phosphorous atom at 3' splice site (3' SS), resulting in the displacement of the 3' end of the intron and joining of the 5' and 3' exons. B: SL-addition *trans*-splicing is very similar to the canonical *cis*-splicing, except for the use of two substrate molecules, which is the reason for the generation of a Y branched intermediate instead of a lariat. C: *cis* regulatory elements involved in pre-mRNA splicing. 5' SS, 3' SS, and BP consensus sequences are shown. Polypyrimidine tract (PPT) and exonic splicing enhancers (ESE) are indicated. D: schematic representation of *trans*-splicing between (a) ABP and (b) HDC pre-mRNAs; (c) represent the mature chimeric mRNA.

resulted negative (Dandekar & Sibbald 1990). Experiments in mammalian cells showed that SL sequences from nematodes and *Leptomonas collosoma* were accurately *trans*-spliced to both nematode and adenovirus suggesting that the machinery utilized in *cis* could perform a *trans*-splicing, therefore both machineries should present common features (Bruzik & Maniatis 1992).

The natural occurrence of *trans*-splicing in mammalian cells was first time proposed for genes encoding immunoglobulin heavy chain, c-*myb* and androgen-binding protein (ABP) products (Shimizu et al. 1989, Sullivan et al. 1991, Vellard et al. 1991). For example, ABP is an androgen carrier protein produced in the testicular Sertoli cell but is also transiently expressed during rat liver development. The analysis of transcripts expressed in rat fetal liver showed existence of hybrid molecules between ABP and histidine decarboxylase (HDC) mRNAs (Fig. 1D). Then, it was proposed that *trans*-splicing generated the hybrid molecules, since the genes encoding these two proteins are localized in different chromosomes. Moreover, the joining sequence between the two mRNAs occurred at nucleotides preceding a canonical donor site (/GU) at the fifth ABP intron and following acceptor (AG/) site of the HDC intron 1 (Fig. 1D). Although these experiments did not address the mechanisms whereby the mammalian hybrid pre-mRNA arises, it suggests that *trans*-splicing could be the responsible one.

At the present time, the pre-mRNA *trans*-splicing in eukaryotic cells is more often detected, suggesting its participation in distinct biological processes. It is important to highlight that in early branched eukaryotes *trans*-splicing is responsible for the joining of a leader sequence that is not translatable, while in mammalian cells *trans*-splicing could generate a different protein product, enhancing protein diversity in the cell. Thus, it is expected that these differences might have different biological meanings.

*Trans*-*splicing of viral transcripts: cryptic sites impacting cell viability?* - The analysis of the transformation potential of different SV40 fragments in rat cells provided the first evidences that mammalian cells were able to join two independently transcribed viral pre-mRNA molecules by *trans*-splicing. SV40 genome is divided into early and late regions, the expression of the early region is required for the induction and maintenance of the transformed cell state. This region encodes the large T and the small t antigens, which are generated by alternative "*cis*-splicing". Two different 5' donor sites and one 3' acceptor site are used for the production of the two mature mRNAs.

Microinjection experiments in rat cells using a plasmid construct bearing the distal part of the large T intron, the small t intron and the second large T exon under the control of the early SV40 promoter resulted in the production of two mature mRNAs (Eul et al. 1995). One of them is the expected mRNA (T2 mRNA), processed by *cis*-splicing with the excision of the small t intron. The other transcript detected (T1 mRNA) is the result of the use of the canonical cryptic 5' donor splice site, located in the second large T exon to join the small t exon using its 3' acceptor site. Since the 3' acceptor site precedes the 5' donor

site, the only possibility for the generation of T1 transcripts is a *trans*-splicing reaction. Interestingly, T1 and T2 mRNAs could be translated producing two distinct antigens, which in turn suggests a possible biological function for the process.

Subsequently the spectrum of virus transcripts that could be processed by *trans*-splicing experiments was broadened, now indicating the joining of two distinct molecules. Mammalian cells microinjected with in vitro produced pre-mRNAs from HIV-nef gene and SV40 T antigen were able to produce hybrid mRNAs (HIV-nef/T-antigen) as well as fusion proteins (Caudevilla et al. 2001b). Both 5' cryptic splice sites of the HIV-nef mRNA were spliced to the 3' acceptor site common to the large T and small t antigens. Moreover, for the first time, it was possible to detect hybrid molecules between viral (HIV-nef mRNA) and cellular transcripts, generated by *trans*-splicing. Eight different HIV-nef/cellular hybrid mRNAs were detected. In four of those products, the HIV-nef transcript contributed with one of their cryptic donor sites while cellular transcripts contributed with the acceptor site (Caudevilla et al. 2001b).

Although the above data demonstrate that mammalian cells can utilize viral transcripts as *trans*-splicing substrates in generating hybrid protein, the biological significance of these products was not addressed as well as its occurrence during the course of an infection. However, the impact of viral infection on host RNA splicing could be evaluated by the detection of a great variety of cellular transcripts, including glyceraldehyde-3-phosphate dehydrogenase (GADH) and β-actin, and exon 2 of the major late transcript of adenovirus (MLT), in the course of the infection (Kikumori et al. 2002). A construction containing the 3' acceptor site of MLT intron 1, a c-*myc* epitope, a polyhistidine tag, and a polyadenylation signal of bovine growth hormone was transfected in mammalian cells. *Trans*-splicing of the construct transcript with cellular RNAs could generate tagged proteins with a third of possibilities because of the start codon frame. The tagged proteins were in fact detected by immunoprecipitation in higher levels than controls, but no apparent specificity was observed, suggesting that *trans*-splicing mediated by the 3' acceptor site of MLT intron 1 could be unregulated. Although these experiments failed to find a specific protein related to cell viability, the use of a tetracycline-inducible system showed that this promiscuous *trans*-splicing has a discrete but significant impact on cell growth (Kikumori et al. 2002).

In the examples mentioned above, the cryptic sites constitute intrinsic characteristics of the genes that lead them to participate in *trans*-splicing process, and are not associated with point mutations of the regular splice sites (Caudevilla et al. 2001b). Since the number of known viral cryptic splice sites is still low, it is unclear if there are other attributes which direct these genes to engage in *trans*-splicing mechanisms.

While the steps of heterologous *trans*-splicing reactions during viral infection are well established, its biological meaning is just being defined. In fact, the *trans*-splicing could represent a transgression in cellular RNA processing affecting the cell growth, and a consequent

mechanism for adenoviral-mediated cell death (Kikumori et al. 2002). It is not clear if specific cellular mRNAs could act as targets of *trans*-splicing producing hybrid proteins involved in cell growth and death (Caudevilla et al. 2001b), or if the effect in cell viability may be non-specific, and related to the interference in the overall regular *cis*-splicing in the cell (Kikumori et al. 2002).

*Mammalian interchromosomal* trans-*splicing: absence of consensus splicing sites* - Genes localized at distinct chromosomes can be the templates for the transcription of two pre-mRNAs engaged in the interchromosomal *trans*-splicing, and as described above for the ABP gene (Sullivan et al. 1991), other examples of mammalian interchromosomal *trans*-splicing have been proposed.

Hybrid mRNAs have been detected among transcripts of the genes encoding human calcium/calmodulin dependent protein kinase II and signal recognition particle 72, respectively on chromosomes 10q22 and 18 (Breen & Ashcroft 1997); human acyl-Coa:cholesterol acyltransferase-1 (ACAT-1) and Xa exon, respectively on chromosomes 1 and 7 (Li et al. 1999); rat leukocyte common antigen-related (LAR) tyrosine phosphatase receptor and a 3' UTR on chromosomes 5 and 1 respectively (Zhang et al. 2003), and the mouse meiotic recombination gene *Msh4* localized on chromosome 3 and three different sequences on chromosomes 16, 2, and 10 (Hirano & Noda 2004).

It is noteworthy that GU-AG rule is not obeyed most of the proposed interchromosomal *trans*-splicing. In fact, the chimeric *ABP-HDC* transcript is the unique example in which the GU-AG rule could be used for the proposition of a conventional *trans*-splicing processing. For example, the 5' and 3' intron boundaries proposed for the generation of β, δ and ε variants of the *Msh4* gene are respectively UG-GU, AU-AU, and UC-CA (Hirano & Noda 2004). However, it is interesting to note that the δ splice donor and acceptor sites are similar to the AU-AC introns.

The almost absence of conventional splicing boundaries in the generation of hybrid mRNA molecules coded by genes located in different chromosomes suggests an alternative mechanism of RNA processing for their generation. However, *trans*-splicing cannot be completely excluded if it is assumed that it occurs through a non-conventional splicing mechanism. So, the study of mechanisms involved in the generation of such hybrid molecules is important for the distinction between a splicing related and other type of RNA processing.

Hybrid mRNAs transcribed from genes in different chromosomes have also been detected in chicken and rice seeds (Vellard et al. 1991, Kawasaki et al. 1999). In chicken, hybrid mRNAs are formed between the transcript of c-*myb* proto-oncogene localized on chromosome 3 and an exon on chromosome 17 (Vellard et al. 1991). In rice, calcium dependent seed-specific protein kinase mRNA is coded by two regions located on chromosomes 6 and 10 (Kawasaki et al. 1999).

In humans, we will highlight the case of ACAT-1. This is a relevant protein in lipid metabolism. Two isoforms of the protein were detected in in vitro expression studies (Yang et al. 2004). One of those isoforms is encoded by a *trans*-spliced mRNA, in which the initiation codon is GGC (Gly) instead of AUG (Met), suggesting a possible biological function related to its diversity. However, ACAT-1 mRNA is rare and the protein produced from the *trans*-spliced RNA could be detected only in macrophages stimulated with phorbol esters (PMA) and in human monocyte-derived macrophages. Moreover, the activity of the *trans*-spliced produced protein is approximately 30% of the other isoform. As the abundant isoform forms tetramers, the possible formation of heterotetramers could constitute a negative regulatory mechanism.

*Intergenic* trans-*splicing of closely linked genes in mammals: lower frequency or a noise reaction?* - Another possibility in the generation of hybrid mRNA molecules is *trans*-splicing of transcripts coded by a cluster of genes. Hybrid mRNAs were detected between members of the immunoglobulin locus (Shimizu et al. 1989, Fujieda et al. 1996), the human GTPase *RSG12* gene and a sequence localized 170 kb downstream from the *RGS12* gene (Chatterjee & Fisher 2000), transcripts coded by genes of the cytochrome P450 3A cluster (Finta & Zaphiropoulos 2002), and members of the mouse protocadherin locus (Tasic et al. 2002).

The four cytochrome P450 3A genes (*CYP3A4*, *CYP3A5*, *CYP3A7,* and *CYPA43*) are located in a cluster of human chromosome 7 (Finta & Zaphiropoulos 2002). *CYPA43* is in a head-to-head orientation to the *CYP3A4* and *CYP3A5* genes, i.e. they are transcribed from different DNA strands. RT-PCR and RNase protection studies showed hybrid molecules between the first exon of *CYPA43* and *CYP3A4* or *CYP3A5* exons. The joining of intergenic exons occurs at canonical splice sites and generates translatable mRNAs, however endogenous protein products were not detected. Moreover, quantitative RT-PCR experiments showed that the expression levels of hybrid molecules is 650:1 in relation to the canonical the mRNAs.

Mammalian protocadherins (Pcdh) are a family of cell surface proteins that could enhance neuronal protein diversity. *Pcdh* gene families (α, β, and γ) are clustered in the genome. The three clusters are localized in a 900 kb region of mouse chromosome 18 (Fig. 2A). Mouse *Pcdh* α contains 14 variable exons while *Pcdh* γ contains 22 variable exons. Each variable (V) exon encodes chaderin extracellular domais, the transmembrane portion and a small piece of the cytoplasmic domain. The three constant exons are located at the 3' end of each cluster and encode the C-terminal part of the cytoplasmic domain. *Pcdh* pre-mRNAs are processed by splicing of one V exon to the first constant exon of each cluster. Otherwise, the mouse *Pcdh* β cluster contains 22 V exons and do not have a constant exon that in theory could encode a cytoplasmic domain, i.e. each single exon encodes a *Pcdh* β protein. It was shown that each α and γ V exon is transcribed through its own promoter, and the promoter choice determines the exon which will be spliced to the first constant exon (Tasic et al. 2002). The α and γ isoforms of Pcdh are generated by alternative *cis*-splicing within a gene cluster, and interchromosomal and intracluster *trans*-splicing was excluded (Tasic et al. 2002). However, intercluster *trans*-splicing could be detected between α or γ V exons and the first

constant exon of the γ or α respectively, as well as β exons and exons from a nearby gene *mDia1* can *trans*-splice to the first constant exon of α or γ clusters (Tasic et al. 2002) (Fig. 2B). This finding suggests the possibility of increase in diversity of Pcdh proteins by intercluster *trans*-splicing, although the levels of *trans*-spliced mRNAs are two orders of magnitude lower than the *cis*-spliced mRNAs, and then functional relevance has to be demonstrated.
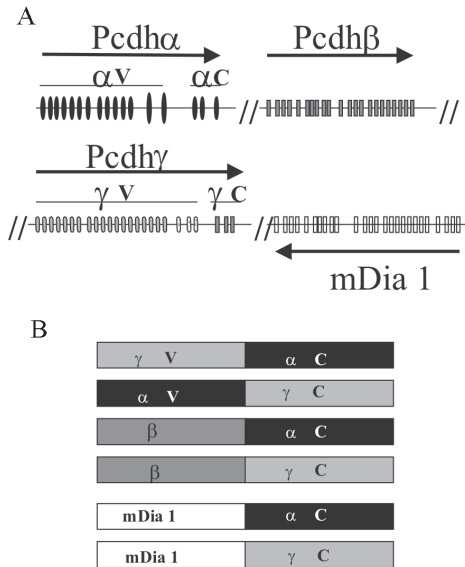


Fig. 2: *trans*-splicing of closely related genes. A: schematic representation of the mouse protocadherin locus. Pcdhα, Pcdhβ, and Pcdhγ indicate the gene clusters. mDia 1 is the neighbor gene. Arrows indicate the transcription direction of each cluster. Variable and constant exons are represented as V and C bars. B: possible intercluster *trans*-spliced mature mRNAs are represented by the joined boxes.

The correlation between promoter activity and the choice of the exon that will be the first in the mature mRNA could be interpreted as an interaction between transcription and splicing (the coupling model) with the recruitment of the splicing machinery to the capped proximal exon (Tasic et al. 2002).

The analysis of protocadherin transcripts showed that only the capped-proximal V exons could be *cis*-spliced or *trans*-spliced to the first constant exon (Tasic et al. 2002). For a *trans*-spliced product, the coupling model cannot be applied since in *trans*-splicing reactions, both pre-mRNAs are independently transcribed. In *trans*-splicing reactions the 5' splice site must pair with 3' splice site from another molecule, and according to the coupling model it would be possible that capped-proximal 5' splice site transiently dissociates from RNA polymerase CTD domain before the synthesis of the 3' splice site, rendering 5' splice site available for paring with another free 3' splice site. In this case, coupling transcription and splicing could minimize inappropriate *trans*-splicing, explaining the low levels of *trans*-spliced mRNAs, and suggesting that *trans*-splicing could be a noise in *cis*-splicing reactions. This proposal is mainly supported by two facts: the low frequency of *trans*-splicing products and the absence of a biological function assigned for them.

*Mammalian* trans-*splicing and exon repetition: efficient but non-essential* - Exon repetition is a term that describes the presence of more than one copy of an exon in mRNAs without alterations at the DNA level. It was proposed that *trans*-splicing between two identical pre-mRNA molecules could be involved since the joining of duplicated exons are precise.

Exon repetition was first described during expression studies of two rat genes, carnitine octanoyl transferase (COT) and medium chain acyl-CoA synthetase (SA) (Caudevilla et al. 1998, Frantz et al. 1999). After that, the observation was made in a small number of other rat and human genes including the rat sensory neuron specific (SNS) voltage-gated sodium channel (Akopian et al. 1999), the human and rat Sp1 transcription factor (Takahara et al. 2000, 2002) and the hERα human estrogen receptor-α (Flouriot et al. 2002) genes.

The COT gene is composed of 17 exons. During studies of COT gene expression in Sprague-Dawley rat liver, two cDNAs with exon duplications were obtained, i.e. instead of the canonical organization exon1-exon2-exon3…exon17, these cDNAs presented the alternative structures exon1-exon2-exon2-exon3…exon17 or exon1-exon2-exon3-exon2-exon3…exon17 (Akopian et al. 1999). An *in silico* analysis of putative protein synthesized from the canonical *cis*-spliced form showed a 70 kDa protein, but in relation to the exon repeated transcripts, only the longest *trans*-spliced alternative form (exon2-exon3-exon2-exon3) is in frame to be translated in a larger (80 kDa) protein. In fact, COT specific antibodies detected two proteins with apparent molecular mass of 69 and 79 kDa, in Western blot experiments for peroxisomal proteins, indicating a putative protein product translated from a possible intragenic *trans*-spliced mRNA. Remarkably, the analysis of COT gene exon repetition in mammalian related species showed it is not conserved, demonstrating that it is not essential, at least for the studied species. The absence of exon repetition in these mammals could be explained by the loss of an ESE sequence in the exon 2, which is present in the rat (Caudevilla et al. 2001a).

A more complex study of exon repetition using two rat lines showed that Sa is a gene that is expressed in the liver and kidneys of rats. The transcript is more abundant in spontaneously hypersensitive rats (SHR) than the normotensive Wistar-Kyoto rats (WKY), although exon repetition was only detected in WKY rats. The observed pattern was exon1-exon2-exon2-exon3-exon4-exon5-exon6 and exon1-exon2-exon3-exon4-exon2-exon3-exon4-exon5-exon6 (Frantz et al. 1999). Remarkably, relative frequency of exon repeated transcripts is as abundant as the canonical ones showing an efficient production of the *trans*-spliced molecule (Rigatti et al. 2004). The concomitant analysis of COT exon 2 nucleotide sequences from these two rat lines showed that both sequences were identical. Moreover, as for the *Sa* transcripts, COT exon 2 repetition was observed only in WKY rats, raising the possibility that *trans*-splicing could be determined by a *trans*-acting

factor and not by a ESE. However, segregation studies using SHR and WKY rats showed that exon repetition is restricted to specific alleles (Rigatti et al. 2004), i.e. the determinant factor for COT transcript exon repetition is a *cis*-regulatory element, but not related to the proposed ESE sequence.

A possible biological role for exon repetition is related to proteome diversity. The only putative product detected is the 80 kDa protein coded by exon1-exon2-exon3-exon2-exon3-exon4-exon5-exon6 COT mRNA. However, this protein is not essential, since exon repetition was not observed in all rat lines studied (Rigatti et al. 2004). Additionally, the COT transcript bearing only the exon 2 repetition is out of frame. It was also observed that a construct bearing exon 2 repetition dropped the expression of the reporter gene, a fact that was interpreted by the authors as a detrimental effect of the exon repetition in natural COT expression (Rigatti et al. 2004). On the other hand, a negative regulation could also be proposed to explain the biological meaning of this exon repetition.

In the absence of other studies related to the detection and function of protein products coded from transcripts bearing repeated exons, it could be argued that the generation of these transcripts could represent a noise reaction. On the contrary, a hallmark of the transcripts bearing repeated exons, besides the low number of genes studied, is their high level of expression (Caudevilla et al. 1998, Rigatti et al. 2004), suggesting a limited but efficient process.

Trans-*splicing in* Drosophila: *efficient and essential* - In contrast to the non-essential *trans*-splicing described for mammals, studies of two *Drosophila loci* showed that the phenomenon generate diversity in the production of two essential proteins (Dorn et al. 2001, Labrador et al. 2001, Horiuchi et al. 2003).

The *Drosophila modifier of mdg4* [*mod(mdg4)*] locus encodes a large number of proteins with different functions especially related with the formation of chromatin complexes (Dorn et al. 2001, Labrador et al. 2001). The observed diversity could be explained by a combination of common first four exons and 26 different terminal exons. Upstream sequences of the fifth exons present AG dinucleotides at the 3' introns boundary and the putative branch point and polypyrimidine tract. As seven out of 26 different 3' ends were encoded on the opposite strand, it was proposed that *trans*-splicing was responsible for the mRNA diversity (Dorn et al. 2001). Fifth 3' exon originated from each opposite or same strand were inserted in a different chromosome of the endogenous locus, to construct two distinct transgenic flies. RNA processing analysis of those flies confirmed that both types of exons were joined to the first four exons, suggesting that *trans*-splicing could account for the generation of all *mod(mdg4)* isoforms. Similarly, when the first four exons bearing its own promoter were expressed on a distinct chromosome, a mutation in the corresponding endogenous sequence could be rescued. All observation in this artificial system showed that chromosomal context is not important for the generation *mod(mdg4)* mature mRNAs. Moreover, multiple TATA-box-containing elements were found throughout

the entire locus, and at least one of the fifth exons that were transcribed from the same strand of the common exons had its own promoter function determined. The results suggested the existence of independent transcriptional units for each fifth exon, although the other 25 promoters were not characterized.

The *Drosophila longitudinals lacking* (*lola*) complex gene (Horiuchi et al. 2003) encodes at least 20 isoforms of BTB-Zn finger transcription factors required for axon guidance decisions in the *Drosophila* nervous system development. The protein is translated from a mRNA variant that contains a constant region (C) composed of exons 5-8, which encodes a N-terminal BTB dimerization domain followed by one or two variable exons (V), which encodes the C-terminal zinc finger variable domain. To complicate the scenario, exons 1 to 4 are alternatively used as the initial 5' exon through an alternative transcription initiation from four possible sites. So, the complex locus spanning 60 kb, where 32 exons are aligned on the same DNA strand, can generate 80 splicing variants by both alternative transcription initiation and alternative splicing. Alternative splicing of the following variable exons (9-32) produces 20 different combinations, 17 of them have unique zinc finger motifs, suggesting different target DNA sequences. The isoforms have a complex pattern of expression, and expression of different isoforms in the same cell is observed, multiplying functional diversity once dimerization of different isoforms could occur.

Mutations where a specific *lola* isoform was inactivated correlate to a specific defect in axon guidance choice point, suggesting that alternative splicing in *lola* locus is important in the determination of axon trajectories. Lethal mutations in constant exons could complement mutations in variable exons localized in different homologous chromosomes. RT-PCR experiments showed that wild type chimeric mRNAs were detected in F1 flies heterozygotes for mutations in V and C exons. Moreover, the joining between the V and C exons synthesized by the two homologous chromosomes occurred at canonical splice-sites (Horiuchi et al. 2003).

It is noteworthy that approximately half of the mature mRNA for some isoforms is originated by *trans*-splicing between pre-mRNAs transcribed from two homologous chromosomes. Also, it was found that at least one of the variable *trans*-spliced exons is probably transcribed from its own promoter. A negative position effect for the frequency of *trans*-splicing was observed in flies bearing an inverted locus with respect to its homologous. This observation, more pronounced if the locus was positioned in a distinct chromosome, suggests that chromosome pairing during transcription, a typical feature in insects, is important for the generation of *trans*-spliced *lola* transcripts.

The position interference observed in *trans*-splicing of the *lola* transcripts is not observed in *modifier of mdg4* [*mod(mdg4)*] locus, since *trans*-splicing of *modifier of mdg4* [*mod(mdg4)*] pre-mRNAs occurred when the transcripts were expressed from different chromosomes (Dorn et al. 2001). However, the expression system of the transgene, GAL4-UAS system, does not represent the natural situation, and different levels of transgene tran-

script could be produced which escape a natural regulation (Horiuchi et al. 2003).

*SL* trans-*splicing: a constitutive reaction* - The classical *trans*-splicing consists in the addition of a short SL sequence to the 5' UTR of pre-mRNAs. Although described at first in Kinetoplastida, SL *trans*-splicing was subsequently shown to occur in other protists of the Euglenozoa phylum, i.e. Euglenida (Tessier et al. 1991, Ebel et al. 1999, Frantz et al. 2000), Diplonemida (Sturm et al. 2001), and Kinetoplastida (Murphy et al. 1986, Sutton & Boothroyd 1986, Laird et al. 1987). In the metazoans, *trans*-splicing has been described in free-living or parasitic nematodes (Krause & Hirsh 1987, Blaxter & Liu 1996), in trematodes (Rajkovic et al. 1990, Davis et al. 1994) and cestodes (Brehm et al. 2000) as well as in turbelarians (Davis 1997) of the Platyhelminthes phylum. More recently, this form of RNA processing has been described in *Hydra*, a member of Cnidaria phylum (Stover & Steele 2001) and, surprisingly in two members of the Urochordata subphylum of chordates, the ascidian *Ciona intestinalis* (Vandenberghe et al. 2001) and the appendiculariam *Oikopleura dioica* (Ganot et al. 2004). It is absent in a variety of organisms where EST libraries were intensively sequenced, e.g. arthropods, plants, most protists, fungi, and vertebrates. This wide but sporadic range of phylogenetic distribution among eukaryotes prompted the question if SL *trans*-splicing arose independently many times or if it was originated once and lost from various lineages (Nilsen 2001, Blumenthal 2004). At present, based on phylogenetic data it is impossible to discriminate between both hypotheses. It is important to stress here that determining a representative phylogenetic distribution of SL *trans*-splicing in eukaryotes is not an easy task because there is little sequence conservation among the identified SL RNAs from the various phyla. A directed EST project intended to cover a wide phylogenetic range could in theory help in the determination of SL *trans*-splicing distribution (Nilsen 2001).

The poor sequence conservation in SL RNAs across the diverse phyla is accompanied by a little length conservation, although they are invariably small RNAs (< 150 nt) (Nilsen 2001). The mini-exon sequence, which is transferred to the 5' end of pre-mRNAs, is also variable in length and ranges from 16 nt to 51 nt (Davis 1997, Vandenberghe et al. 2001). Most of the SL RNAs showed a variable secondary structure. Nevertheless, a common secondary structure composed of three stem-loops is observed in kinetoplastids and nematodes (Nilsen 2001).

SL *trans*-splicing and *cis*-splicing are very similar and little differences were observed to date. Intron boundaries are defined by the same elements observed in *cis*-splicing, although the 5' donor splice site is localized in the SL RNA molecule and the 3' acceptor splice site is localized in the pre-mRNAs. The 5' intron boundary is defined by a GU dinucleotide, while the 3' intron boundary is defined by an AG dinucleotide (see Fig. 1). A polypyrimide tract is observed upstream of the AG dinucleotides in kinetoplastids (Liang et al. 2003), cnidarians and in *C. intestinalis* (Vandenberghe et al. 2001), whereas in nematodes there is no polypyrimidine tract associated

to the 3' splice site. In nematodes a conserved sequence UUUCAG/ (AG, acceptor site) is required for proper processing (Conrad et al. 1993, Romfo et al. 2001). Finally, in both processes it is possible to find out the participation of ESEs, SR proteins as well as similar spliceosome which are formed by almost the same snRNPs and proteins (Sanford & Bruzik 1999, Liang et al. 2003).

The difference between the two types of processing resides in the components of the 5' splice site recognition. While U1 snRNP is required in *cis*-splicing, it does not participate in SL *trans*-splicing. In fact, the SL RNA acts as a substrate as well as a component of the spliceosome catalytic complex (Liang et al. 2003). Moreover, in nematodes the SL snRNP contains two proteins that are exclusive to the SL snRNP of nematodes and essential to the SL *trans*-splicing but not for *cis*-splicing (Denker et al. 2002).

In kinetoplastids SL *trans*-splicing processes virtually all pre-mRNAs, while, only one gene was found to have its mRNA processed by *cis*-splicing (Mair et al. 2000). This prevalence raised the question of the biological role of *trans*-splicing in those organisms. Kinetoplastids present polycistronic transcription. Moreover, a transcriptional analysis of two *Leishmania major* chromosomes showed few promoter regions to drive the expression of their protein coding genes (Martinez-Calvillo et al. 2003, Worthey et al. 2003). In *L. major* chromosome 1 the transcription starts at a single strand-switch region and goes bi-directionally towards each telomere (Martinez-Calvillo et al. 2003). An analogous analysis of chromosome 3 showed that transcription starts near each telomere towards a region occupied by a tRNA gene at which transcription terminates (Worthey et al. 2003). These observations clearly demonstrate that the transcription units in kinetoplastids are very long and SL *trans*-splicing functions in the individualization of the messages. A single type of SL RNA is *trans*-spliced to every cistron with the peculiarity that polyadenylation is coupled to the splicing process (LeBowitz et al. 1993). In fact, in these organisms a polyadenylation cleavage site is not a consensus but it is determined by a fixed distance in relation to the downstream polypyrimidine tract of neighbor cistron (Matthews et al. 1994). The *trans*-splicing of the downstream gene occurs before the polyadenylation of the upstream sequence (LeBowitz et al. 1993).

Part of nematode *C. elegans* genome is transcribed as polycistronic units and can be related to the prokaryote operons (Blumenthal & Gleason 2003). In prokaryotes polycistronic transcription results in a single multi-gene mRNA, which is translated into distinct proteins on ribosomes. The operon structure allows the co-regulation of genes of the same metabolic pathway. In *C. elegans*, the polycistronic transcript is first processed by *trans*-splicing and then transported to the cytoplasm to be translated. *C. elegans* has two types of SL RNAs involved in *trans*-splicing, namely SL1 and SL2. SL1 is responsible for the processing of the majority of pre-mRNAs engaged in *trans*-splicing, mostly those not organized in operons. For the pre-mRNAs organized in operons, SL1 is *trans*-spliced to the 5' end of the first cistron, i.e. it is *trans*-

spliced to the acceptor site nearest to the promoter. SL2 is then used in *trans*-splicing the following downstream regions, i.e. it is a specialized form of SL whose function is to generate individual mRNAs from a polycistronic precursor. It was demonstrated that approximately 15% of *C. elegans* genes are organized in operons and more than 90% of these genes are SL2 *trans*-spliced (Zorio et al. 1994, Blumenthal et al. 2002). In fact, there is another type of operon, present in relatively small number, in which SL1 is the only form of SL RNA utilized to process the polycistronic transcript (Blumenthal 2004).

The individualization of mature mRNAs involving SL *trans*-splicing in *C. elegans* is very similar to the kinetoplastid processing, although differences exist. In *C. elegans* operons the first step is the polyadenylation of the upstream sequence to which follows the *trans*-splicing of the downstream transcript. The AAUAAA polyadenylation signal is approximately 100 bp apart from the downstream acceptor site, and its destruction decreases dramatically the *trans*-splicing of the downstream gene (Blumenthal 2004).

Operons were found in the trematode *Schistosoma mansoni* (Davis & Hodgson 1997) and recently, in the urochordate *O. dioica* (Ganot et al. 2004). The operon organization, likewise for *C. elegans*, is related to SL *trans*-splicing. These data suggest that in eukaryotes one strategy for individualizing mRNAs from long polycistronic transcripts is SL *trans*-splicing.

Another important feature of SL RNAs is that they have a 5' cap structure. The addition of $m^7G$ is essential for eukaryotic RNA metabolism and processing. In the nucleus, it is involved in pre-mRNA splicing and in directing mRNAs and UsnRNAs to the cytoplasm. In the cytoplasm, it is related to mRNA stability and translational initiation. So, the transfer of SL sequence to the 5' UTR of pre-mRNAs results in the capping of the mRNA, which persists during the mature mRNA life (Tschudi & Ullu 2002). Thus a SL *trans*-splicing reaction can be considered as a *trans*-capping reaction, making part of the repertoire of biological roles of SL *trans*-splicing.

In kinetoplastids, the $m^7G$ is linked by a 5'-5' triphosphate bridge to the first SL sequence nucleotide (Tschudi & Ullu 2002). The cap structure however is more complex and is named cap 4 because the first four 5' nucleotides adjacent to the cap are modified by methylations. By analogy with other eukaryotic biological systems, the cap 4 structure was always implicated in mRNA stability, transport and translational regulation. A few indirect evidences support this analogy, e.g. mutant pre-mRNAs that do not *trans*-splice efficiently do not accumulate in vivo (Ullu et al. 1996). A mutational analysis of SL sequence showed that the part of the sequence itself and/or cap 4 formation are relevant for the association of mature mRNAs with the polysomes, suggesting the importance of this structure in the translation of kinetoplastid pre-mRNAs (Zeiner et al. 2003). Finally, in vitro analysis of cap 4 binding to recombinant eIF4E translation factor of *Leishmania* demonstrates its association with the translation machinery (Lewdorowicz et al. 2004).

In nematodes SL 1 function is not associated to the individualization of polycistronic transcripts. In *C. elegans*

the predominant form of *trans*-splicing is SL1 *trans*-splicing (57%) (Zorio et al. 1994) as well as for *Ascaris suum* (70%) (Maroney et al. 1995). In *Hydra* (Stover & Steele 2001) and *C. intestinalis* (Vandenberghe et al. 2001) there is no described operon structure to date. Taken altogether, these observations suggested that another role for SL *trans*-splicing exists in these organisms. SL1 and SL2 RNAs of *C. elegans*, and the SL RNAs of *Hydra* and *C. intestinalis* have a common cap structure, the N-2,2,7 trymethylguanosine cap ($m_3^{2,2,7}GppG$ or TMG cap). As it was suggested for kinetoplastids, it was postulated that SL structure and TMG cap could modulate the translational efficiency of mRNAs. Recently, the effect of *trans*-splicing on translation efficiency was assayed in *Ascaris* embryo cell-free translation system showing that each TMG cap or spliced leader sequence alone decreased the activity of a reporter gene (Lall et al. 2004). However, both features act synergistically to promote efficient translation, suggesting that SL *trans*-splicing has a biological role in promoting translation.

So, resolving polycistronic transcripts and promoting translation are two biological functions recognized up to now for SL *trans*-splicing. A more precise map of the phylogenetic distribution of *trans*-splicing, as well as the functional analysis of this type of processing will probably solve the questions of its origin and other putative function(s).

*Alternative SL* trans-*splicing: variability or noise?* - The majority of SL *trans*-splicing events so far described in kinetoplastids pointed to a precise use of 3' acceptor sites. However, many trials have shown that more than one acceptor site within the same intergenic region can be used in the generation of mature mRNAS coding the same ORF but distinct 5' UTR extensions (Vassella et al. 1994, Nepomuceno-Silva et al. 2001, da Silva et al. 2002, Manning-Cela et al. 2002).

Expression studies of the single copy arginase gene in the kinetoplastid *L. (L.) amazonensis* showed that SL was added to more than two different acceptor sites in the pre-mRNA (da Silva et al. 2002) (Fig. 3A). Nucleotide sequence analysis of the longest cDNA demonstrated that two consecutive AG dinucleotides (positions –272 and –270) can be used as acceptor sites, while sequencing of the short cDNA showed that another two acceptor sites (positions –137 and –129) can either be used (da Silva et al. 2002). Thus the two far apart regions can be used as acceptor sites, each one displaying a local micro-heterogeneity. Interestingly, each one of the two acceptor regions is preceded by a polypyrimidine tract, an important element for 3' splice acceptor site determination.

The role for the presence of more than one 3' acceptor site in the arginase gene is uncertain, however it could be related to an evolutionary drive that permits the accumulation of acceptor sites upstream of the ORF, which guarantee the pre-mRNA processing by *trans*-splicing. In kinetoplastids, alterations in acceptor site usage of one gene could implicate in alternative polyadenylation sites of the upstream gene. Since there is no polyadenylation signal sequence and polyadenylation of the upstream gene occurs at a ~100 nt distance subsequently to the

**A**

```
GATCGACATGTTGAACAAAAACGTCTCTCCCGTCTCATCTGCTCCACTTGTATGCGATGACTTTACACAC
ACACATATACGTATAACGTGTACGTGTACGTCTCCCCACCTTCCCCCAACGCTGTTATCCTGCCCTGCTC
CTCCTCGCAGAGAATTAAAATTTGCCCCTTGAAGCAACGACACGACAAGGTAAAACACCGTAGTGCTCCT
CCCTTCCCTTCTCTTCTGTTCTGTTCCCTTCCCTCACACGTTATTCATTGCATTCAACCGAAATTTGCTT
TCAAGCGAAACAGGAAAACGCAACCCCCGCATACGAACCAAGAAAAACAAAAAAGGAGGCAGGTCTGCCG
TTTGTCATCCAGTCTCACCGGCACCCACCCCACCGATATCCGCAGCCAACCCCCACCTCCTCCCGTTCAT
CATGGAGCACGTGCAGCAGTACAAGTTCTACAAGGAGAAGAAGATGAGCATTGTGCTTG
```

**B**

```
GCTTCGCAAC TTGCCTTTTT GCTCTCTGCC TTGTTGAAGT TTCCCCTCGT TTTTTTTTTT
TTGTTTTG TGGCCGTGCT CGCCCTCATG CGGAAGAAAG CCGCAGCATT AGTAGCACCCAC
AGCAGACA CACGGCCGAC GTGCCGCGGG GCTGCCATTG CGAATAACTT TATGGGACGTGC
CGGCCCCC G
```

**C**

```
CACGTACTGGTCTTTCTCTTTTTTCCTATTTTTCTTTGCACGTTCTTTGCCTCCGCGCCATGCGGCCCAC
CCTTCCTCTCCCCTCCCCTGCACGCCCGCGAATGCGCGATCGACACCACGGCCAAACCAATAATGCTTGC
CGCATCTTCTACGCCGCTCACGCCCGCCACCACCACCGGCAGTACCGCCGTTTTAGTGCTACAACAACAA
                A              T  C         G
CAAAGTGTTTCGTACCCTTTTCTTCGTTTCTTCCCCGGTACACCCCCGCGCACGCGAACGGCTCCGTGAG...
GTGCCGCCCGCATCCTTGCAAGCGCCGACGATGAGCTCCAAAAAG
```

Fig. 3: alternative *trans*-splicing in trypanosomatids. A: nucleotide sequence of arginase form *Leishmania (L.) amazonensis* (GenBank accession no. AF038409). Boxes indicate the polypyrimidine tract, AG 3' acceptor site is indicated in gray, bold characters indicate the beginning of the ORF. B: nucleotide sequence of *LYT1* of *Trypanosoma cruzi* (GenBank accession no. AF263616). Symbols are the same as A, with the addition of italic bold that representing an alternative initiation codon. C: nucleotide sequence of LDH1A of *L. donovani* (GenBank accession no. AF109296). Symbols are the same as A, the line in italic indicates the nucleotide variations found in the allele.

*trans*-splicing of the downstream transcript. The alternative acceptor sites could be used as a regulatory mechanism for the expression of upstream genes. It is important to highlight here that gene expression in trypanosomatids are mainly post-transcriptionally regulated, and different 5' UTR and 3' UTR extensions could contribute with this regulation by generating different UTR structures and/or different targets for the binding of factors.

Alternative acceptor splice sites were also found for *Trypanosoma cruzi TcRho1* and *Lyt1* transcripts (Nepomuceno-Silva et al. 2001). TcRho 1 is a GTPase member of the Ras superfamily, which are involved in diverse signal transduction pathways. The analysis of 5' UTR of *TcRho1* transcripts showed that five alternative AG splice acceptor sites are used to generate five different mature mRNAs. Two acceptor sites were mapped in the intergenic region but the other three were found within the ORF. The alternative ORFs downstream of the three TcRho1 ORF internal acceptor sites could direct the synthesis of putative small protein 54 residues long. Nevertheless, no similar proteins were detected in protein databanks. Two different polypyrimidine tracts were detected, one in a region localized upstream of the intergenic acceptor sites while a short polypyrimidine tract is present upstream of the three

ORF internal acceptor sites. As proposed for the *Leishmania* arginase gene, a similar biological role could be attributed for the intergenic acceptor sites in *T. cruzi* TcRho1 gene. On the other hand the utilization of acceptor sites inside ORFs could contribute with a negative post-transcriptional regulation of gene expression since it would generate truncated mRNA (Nepomuceno-Silva et al. 2001).

An alternative explanation for the biological meaning of acceptor sites inside ORFs emerged from studies of the LYT1 transcripts in *T. cruzi*. LYT1 protein (LYT1p) participates in diverse, apparently non-related, biological processes (Manning-Cela et al. 2002). Alternative *trans*-splicing produced three different transcripts; two of them were processed in the intergenic region while a third type was processed inside the ORF (Fig. 3B). One of the transcripts processed in the intergenic region is present in lower amounts than the others and its acceptor site is a non-canonical GG dinucleotides. The ORF analysis showed a putative signal sequence at its N-terminal part and a nuclear localization sequence at its C-terminal moieties. The alternative addressing of LYT1p to the nucleus or to the cell surface could explain the diverse roles attributed to this protein. It is interesting to note that the transcript

processed inside the ORF would be translated in a truncated protein that lacks the signal sequence. Moreover, the amount of the two major transcripts is regulated through the parasite development. This way the alternative *trans*-splicing generates two transcripts that could be translated in two different proteins with apparent different localization in the cell and participating in different processes (Manning-Cela et al. 2002).

Alternative SL-*trans*-splicing in kinetoplastids is a mechanism able to produce different transcripts from the same pre-mRNA. Otherwise, the utilization of non-canonical acceptor sites could simply reflect a noise in the major reaction.

A different splice acceptor site was mapped in the intergenic region of the proton translocating P-type adenosine triphosphatase LDH1A of *L. donovani* and was pointed to be a result of an allelic variation, but it only occurred when the cells were maintained for a high number of culture passages (Fig. 3C). So, care should be taken in relation to allelic variations and/or variant appearance in long term culture stocks (Stiles et al. 1999). Interestingly, these results correlate with the observation, in *T. cruzi*, of a mutation that resulted in the loss of the canonical AG and lead to the use of the next AG dinucleotide as an acceptor (Hummel et al. 2000).

It is interesting to note that *cis* and *trans*-splicing occur in the same nuclear environment. A tight regulation was shown in *C. elegans* where *cis*-splice donor sites are dominant over SL-donor site. However, *trans*-splicing at *cis*-acceptor sites have been described in trypanosomes, trematodes and cestodes, as recently discussed by Hastings (2005). Those observations could represent another way of generating transcript variability.

As a conclusion, splicing is an important regulatory mechanism in the development of organisms and also a key to assure genetic variability. Post genomic information has been showing that the number of expected genes in many organisms is lower than the estimated. Part of the variability is generated at the level of RNA molecules. Alternative *trans*-splicing of transcripts could contribute to the expansion of protein diversity by raising the possibilities of exon combinations.

In mammals, the features considered to engage two independently transcribed pre-mRNA molecules in *trans*-splicing are the absence of DNA recombination and the joining of RNA sequences through the use of putative canonical splice sites. Interchromosomal *trans*-splicing is the most difficult processing to demonstrate, since just one of the described examples has GU-AG intron boundaries at putative splice sites, and although a non-conventional splicing processing could explain the produced molecule, another type of RNA processing could not be ruled out. Processing of independently transcribed RNA from closely linked genes, that obey the minimal features to be considered as *trans*-splicing, presents lower frequencies when compared to the *cis*-splicing counterpart, but the protein products translated from the *trans*-spliced molecules have not been detected. The lower frequencies and the scarce data of proteins produced by the *trans*-spliced mRNAs lead to the hypothesis that the phenomenon could be a noise in the *cis*-splicing of two different pre-mRNAs, transcribed in distinct sites, and eventually paired. This interpretation could explain the high frequencies of intragenic *trans*-splicing, since it is expected that splice sites in transcripts generated by the same factory have a higher probability of interaction. However, this hypothesis does not explain why intragenic *trans*-splicing is so uncommon or why, at least in one case, exon repetition is allele-specific. In this particular case, *cis* regulatory elements were proved to be important for the processing to occur. Although these elements and their mode of action were not determined, it is possible that they are restricted to a low number of alleles and, could be related to specific transcription and splicing characteristics.

Exon duplication proved to be highly efficient but at least in one case it is not essential. In contrast, *Drosophila trans*-splicing proved to be efficient and essential, i.e. it is a form of expanding protein variability with a functional biological role. Frequencies of *lola* interallelic *trans*-splicing are very high when compared to the mammalian intergenic *trans*-splicing, but a little higher or similar when compared with intragenic *trans*-splicing. This observation suggests an effect of proximity and attests the above-mentioned idea that the closer two transcription and/or splicing factories are the more likely their respective pre-mRNAs are of being *trans*-spliced. Following this hypothesis, we could consider *trans*-splicing as a possible mechanism to generate protein diversity in metazoa and that *Drosophila* consolidated the process by coupling it into a different mode of transcription where the homologous chromosomes are still paired.

In spite of the examples described up to now, it remains difficult to assign a biological function for *trans*-splicing. A clearer scenario occurs for by SL-addition *trans*-splicing studies. One function of SL-addition *trans*-splicing is the individualization of messages from polycistronic mRNAs in a concerted action with polyadenylation. This type of processing is essential in organisms that have genes organized in operons, like *C. elegans*, or have long polycistronic units, like trypanosomatids. This regulatory role is extreme in trypanosomatids where transcription rates of different genes are very similar and transcriptional units may be as large as half of a chromosome. Another function of SL-addition *trans*-splicing is related to the capping of the RNA molecule. Thus it could be considered as a *trans*-capping reaction connected to cytoplasm addressing, mRNA stability and interactions with the translation machinery.

Recently, the discovery of alternative SL-addition *trans*-splicing in trypanosomatids raised the possibility of regulating gene expression by generating mature transcripts with different 5' and 3' UTRs. The impact of alternative *trans*-splicing was extended since alternative SL addition *trans*-splicing could occur inside ORFs resulting in different protein products.

It is worth mentioning that SL-addition *trans*-splicing has been considered as a general mechanism required for the production of functional mature transcripts, while mammalian *trans*-splicing is proposed to be involved in generating protein diversity. It could be the case that different functions have evolved from a common chemical language of two transesterification reactions and then

adjusted to specific modes of gene expression. Some organisms like *Drosophila* and trypanosomatids developed efficient strategies in utilizing *trans*-splicing as a source of gene expression. Eukaryotic viruses, which speak this same language, could have utilized it to interfere in cellular growth and death by refining some language structures, their cryptic sites. However, it is not possible to exclude completely the proposition that mammalian *trans*-splicing and alternative *trans*-splicing of trypanosomatids could represent a noise reaction with no associated biological function. Acquisition of more data is important for the discrimination of *trans*-splicing as simple side reactions from part of the gene expression repertoire that once was a noise but now is a bang for certain organisms.

## REFERENCES

Akopian AN, Okuse K, Souslova V, England S, Ogata N, Wood JN 1999. *Trans*-splicing of a voltage-gated sodium channel is regulated by nerve growth factor. *FEBS Lett 445*: 177-182.

Blaxter M, Liu L 1996. Nematode spliced leaders: ubiquity, evolution and utility. *Int J Parasitol 26*: 1025-1033.

Blumenthal T 2004. Operons in eukaryotes. *Brief Funct Genomic Proteomic 3*: 199-211.

Blumenthal T, Gleason KS 2003. *Caenorhabditis elegans* operons: form and function. *Nat Rev Genet 4*: 112-120.

Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, Kim SK 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature 417*: 851-854.

Boothroyd JC, Cross GA 1982. Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5' end. *Gene 20*: 281-289.

Breen MA, Ashcroft SJ 1997. A truncated isoform of Ca2+/calmodulin-dependent protein kinase II expressed in human islets of Langerhans may result from trans-splicing. *FEBS Lett 409*: 375-379.

Brehm K, Jensen K, Frosch M 2000. mRNA *trans*-splicing in the human parasitic cestode *Echinococcus multilocularis*. *J Biol Chem 275*: 38311-38318.

Bruzik JP, Maniatis T 1992. Spliced leader RNAs from lower eukaryotes are *trans*-spliced in mammalian cells. *Nature 360*: 692-695.

Burge C, Tuschl T, Sharp, PA 1999. Splicing of precursors to mRNAs by the spliceosomes. In RF Gesteland, TR Cech, JF Atkins (eds), *The RNA World*, Cold Spring Harbor Press, New York, p. 525-559.

Caudevilla C, Codony C, Serra D, Plasencia G, Roman R, Graessmann A, Asins G, Bach-Elias M, Hegardt FG 2001a. Localization of an exonic splicing enhancer responsible for mammalian natural *trans*-splicing. *Nucleic Acids Res 29*: 3108-3115.

Caudevilla C, Da Silva-Azevedo L, Berg B, Guhl E, Graessmann M, Graessmann A 2001b. Heterologous HIV-nef mRNA *trans*-splicing: a new principle how mammalian cells generate hybrid mRNA and protein molecules. *FEBS Lett 507*: 269-279.

Caudevilla C, Serra D, Miliar A, Codony C, Asins G, Bach M, Hegardt FG 1998. Natural *trans*-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc Natl Acad Sci USA 95*: 12185-12190.

Chatterjee TK, Fisher RA 2000. Novel alternative splicing and nuclear localization of human RGS12 gene products. *J Biol Chem 275*: 29660-29671.

Conrad R, Liou RF, Blumenthal T 1993. Functional analysis of a *C. elegans trans*-splice acceptor. *Nucleic Acids Res 21*: 913-919.

da Silva ER, Castilho TM, Pioker FC, Tomich de Paula Silva CH, Floeter-Winter LM 2002. Genomic organisation and transcription characterisation of the gene encoding *Leishmania (Leishmania) amazonensis* arginase and its protein structure prediction. *Int J Parasitol 32*: 727-737.

Dandekar T, Sibbald PR 1990. *Trans*-splicing of pre-mRNA is predicted to occur in a wide range of organisms including vertebrates. *Nucleic Acids Res 18*: 4719-4725.

Davis RE 1997. Surprising diversity and distribution of spliced leader RNAs in flatworms. *Mol Biochem Parasitol 87*: 29-48.

Davis RE, Hodgson S 1997. Gene linkage and steady state RNAs suggest *trans*-splicing may be associated with a polycistronic transcript in *Schistosoma mansoni. Mol Biochem Parasitol 89*: 25-39.

Davis RE, Singh H, Botka C, Hardwick C, Ashraf el Meanawy M, Villanueva J 1994. RNA *trans*-splicing in Fasciola hepatica. Identification of a spliced leader (SL) RNA and SL sequences on mRNAs. *J Biol Chem 269*: 20026-20030.

De Lange T, Berkvens TM, Veerman HJ, Frasch AC, Barry JD, Borst P 1984. Comparison of the genes coding for the common 5' terminal sequence of messenger RNAs in three trypanosome species. *Nucleic Acids Res 12*: 4431-4443.

De Lange T, Liu AY, Van der Ploeg LH, Borst P, Tromp MC, Van Boom JH 1983. *Tandem* repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription? *Cell 34*: 891-900.

Denker JA, Zuckerman DM, Maroney PA, Nilsen TW 2002. New components of the spliced leader RNP required for nematode *trans*-splicing. *Nature 417*: 667-670.

Dorn R, Reuter G, Loewendorf A 2001. Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in *Drosophila. Proc Natl Acad Sci USA 98*: 9724-9729.

Ebel C, Frantz C, Paulus F, Imbault P 1999. *Trans*-splicing and *cis*-splicing in the colourless Euglenoid, *Entosiphon sulcatum. Curr Genet 35*: 542-550.

Eul J, Graessmann M, Graessmann A 1995. Experimental evidence for RNA *trans*-splicing in mammalian cells. *EMBO J 14*: 3226-3235.

Finta C, Zaphiropoulos PG 2002. Intergenic mRNA molecules resulting from *trans*-splicing. *J Biol Chem 277*: 5882-5890.

Flouriot G, Brand H, Seraphin B, Gannon F 2002. Natural *trans*-spliced mRNAs are generated from the human estrogen receptor-alpha (hER alpha) gene. *J Biol Chem 277*: 26244-26251.

Frantz C, Ebel C, Paulus F, Imbault P 2000. Characterization of *trans*-splicing in Euglenoids. *Curr Genet 37*: 349-355.

Frantz SA, Thiara AS, Lodwick D, Ng LL, Eperon IC, Samani NJ 1999. Exon repetition in mRNA. *Proc Natl Acad Sci USA 96*: 5400-5405.

Fujieda S, Lin YQ, Saxon A, Zhang K 1996. Multiple types of chimeric germ-line Ig heavy chain transcripts in human B cells: evidence for *trans*-splicing of human Ig RNA. *J Immunol 157*: 3450-3459.

Ganot P, Kallesoe T, Reinhardt R, Chourrout D, Thompson EM 2004. Spliced-leader RNA *trans*-splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol Cell Biol 24*: 7795-7805.

Hastings KEM 2005. SL trans-splicing: easy come or easy go? *Trends Genet 21*: 240-247.

Hirano M, Noda T 2004. Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA. *Gene 342*: 165-177.

Horiuchi T, Giniger E, Aigaki T 2003. Alternative *trans*-splicing of constant and variable exons of a *Drosophila* axon guidance gene, lola. *Genes Dev 17*: 2496-2501.

Hummel HS, Gillespie RD, Swindle J 2000. Mutational analysis of 3' splice site selection during *trans*-splicing. *J Biol Chem 275*: 35522-35531.

Kawasaki T, Okumura S, Kishimoto N, Shimada H, Higo K, Ichikawa N 1999. RNA maturation of the rice SPK gene may involve *trans*-splicing. *Plant J 18*: 625-632.

Kikumori T, Cote GJ, Gagel RF 2002. Naturally occurring heterologous *trans*-splicing of adenovirus RNA with host cellular transcripts during infection. *FEBS Lett 522*: 41-66.

Konarska MM, Padgett RA, Sharp PA 1985. *Trans*-splicing of mRNA precursors in vitro. *Cell 42*: 165-171.

Krause M, Hirsh D 1987. A *trans*-spliced leader sequence on actin mRNA in *C. elegans*. *Cell 49*: 753-761.

Labrador M, Mongelard F, Plata-Rengifo P, Baxter EM, Corces VG, Gerasimova TI 2001. Protein encoding by both DNA strands. *Nature 409*: 1000.

Laird PW, Zomerdijk JC, de Korte D, Borst P 1987. *In vivo* labelling of intermediates in the discontinuous synthesis of mRNAs in *Trypanosoma brucei*. *EMBO J 6*: 1055-1062.

Lall S, Friedman CC, Jankowska-Anyszka M, Stepinski J, Darzynkiewicz E, Davis RE 2004. Contribution of *trans*-splicing, 5' -leader length, cap-poly(A) synergism, and initiation factors to nematode translation in an *Ascaris suum* embryo cell-free system. *J Biol Chem 279*: 45573-45585.

LeBowitz JH, Smith HQ, Rusche L, Beverley SM 1993. Coupling of poly(A) site selection and *trans*-splicing in *Leishmania*. *Genes Dev 7*: 996-1007.

Lewdorowicz M, Yoffe Y, Zuberek J, Jemielity J, Stepinski J, Kierzek R, Stolarski R, Shapira M, Darzynkiewicz E 2004. Chemical synthesis and binding activity of the trypanosomatid cap-4 structure. *RNA 10*: 1469-1478.

Li BL, Li XL, Duan ZJ, Lee O, Lin S, Ma ZM, Chang CC, Yang XY, Park JP, Mohandas TK, Noll W, Chan L, Chang TY 1999. Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes. *J Biol Chem 274*: 11060-11071.

Liang XH, Haritan A, Uliel S, Michaeli S 2003. *Trans* and *cis*-splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell 2*: 830-840.

Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, Falcone FH, Gavrilescu C, Montgomery JL, Santori MI, Stern LS, Wang Z, Ullu E, Tschudi C 2000. A new twist in trypanosome RNA metabolism: *cis*-splicing of pre-mRNA. *RNA 6*: 163-169.

Maniatis T, Tasic B 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature 418*: 236-243.

Manning-Cela R, Gonzalez A, Swindle J 2002. Alternative splicing of LYT1 transcripts in *Trypanosoma cruzi*. *Infect Immun 70*: 4726-47288.

Maroney PA, Denker JA, Darzynkiewicz E Laneve R, Nilsen TW 1995. Most mRNAs in the nematode *Ascaris lumbricoides* are *trans*-spliced: a role for spliced leader addition in translational efficiency. *RNA 1*: 714-723.

Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ 2003. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell 11*: 1291-1299.

Matthews KR, Tschudi C, Ullu E 1994. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev 8*: 491-501.

Milhausen M, Nelson RG, Sather S, Selkirk M, Agabian N 1984. Identification of a small RNA containing the trypanosome spliced leader: a donor of shared 5' sequences of trypanosomatid mRNAs? *Cell 38*: 721-729.

Moore M, Query CC, Sharp PA 1993. Splicing of precursors to mRNA by the spliceosome. In RF Gesteland, JF Atkins (eds), *The RNA World*, Cold Spring Harbor Press, New York, p. 303-357.

Murphy WJ, Watkins KP, Agabian N 1986. Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for *trans*-splicing. *Cell 47*: 517-525.

Nepomuceno-Silva JL, Yokoyama K, de Mello LD, Mendonca SM, Paixao JC, Baron R, Faye JC, Buckner FS, Van Voorhis WC, Gelb MH, Lopes UG 2001. TcRho1, a farnesylated Rho family homologue from *Trypanosoma cruzi*: cloning, *trans*-splicing, and prenylation studies. *J Biol Chem 276*: 29711-29718.

Nilsen TW 2001. Evolutionary origin of SL-addition *trans*-splicing: still an enigma. *Trends Genet 17*: 678-680.

Parsons M, Nelson RG, Watkins KP, Agabian N 1984. Trypanosome mRNAs share a common 5' spliced leader sequence. *Cell 38*: 309-316.

Rajkovic A, Davis RE, Simonsen JN, Rottman FM 1990. A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. *Proc Natl Acad Sci USA 87*: 8879-8883.

Rigatti R, Jia JH, Samani NJ, Eperon IC 2004. Exon repetition: a major pathway for processing mRNA of some genes is allele-specific. *Nucleic Acids Res 32*: 441-446.

Romfo CM, Maroney PA, Wu S, Nilsen TW 2001. 3' splice site recognition in nematode *trans*-splicing involves enhancer-

dependent recruitment of U2 snRNP. *RNA 7*: 785-792.

Sanford JR, Bruzik JP 1999. SR proteins are required for nematode *trans*-splicing in vitro. *RNA 5*: 918-928.

Shimizu A, Nussenzweig MC, Mizuta TR, Leder P, Honjo T 1989. Immunoglobulin double-isotype expression by *trans*-mRNA in a human immunoglobulin transgenic mouse. *Proc Natl Acad Sci USA 86*: 8020-8023.

Solnick D 1985. *Trans*-splicing of mRNA precursors. *Cell 42*: 157-164.

Stiles JK, Hicock PI, Kong L, Xue L, Meade JC 1999. *Leishmania donovani* proton translocating P-type adenosine triphosphatases LDH1A and LDH1B: trans-splicing and polyadenylation of transcripts in amastigotes and promastigotes. *Mol Biochem Parasitol 103*: 105-109.

Stover NA, Steele RE 2001. *Trans*-spliced leader addition to mRNAs in a cnidarian. *Proc Natl Acad Sci USA 98*: 5693-5698.

Sturm NR, Maslov DA, Grisard EC, Campbell DA 2001. *Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida. *J Eukaryot Microbiol 48*: 325-331.

Sullivan PM, Petrusz P, Szpirer C, Joseph DR 1991. Alternative processing of androgen-binding protein RNA transcripts in fetal rat liver. Identification of a transcript formed by *trans*-splicing. *J Biol Chem 266*: 143-154.

Sutton RE, Boothroyd JC 1986. Evidence for *trans*-splicing in trypanosomes. *Cell 47*: 527-535.

Takahara T, Kasahara D, Mori D, Yanagisawa S, Akanuma H 2002. The *trans*-spliced variants of Sp1 mRNA in rat. *Biochem Biophys Res Commun 298*: 156-162.

Takahara T, Kanazu SI, Yanagisawa S, Akanuma H 2000. Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic *trans*-splicing. *J Biol Chem 275*: 38067-38072.

Tasic B, Nabholz CE, Baldwin KK, Kim Y, Rueckert EH, Ribich SA, Cramer P, Wu Q, Axel R, Maniatis T 2002. Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol Cell 10*: 21-33.

Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P 1991. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. *EMBO J 10*: 2621-2625.

Tschudi C, Ullu E 2002. Unconventional rules of small nuclear RNA transcription and cap modification in trypanoso-

matids. *Gene Expr 10*: 3-16.

Ullu E, Tschudi C, Gunzl A 1996. *Trans*-splicing in trypanosomaid protozoa. In DF Smith, M Parsons (eds), *Molecular Biology of Parasitic Protozoa,* IRL Press, Oxford, p.115-133.

Vandenberghe AE, Meedel TH, Hastings KE 2001. mRNA 5'-leader trans-splicing in the chordates. *Genes Dev 15*: 294-303.

Vassella E, Braun R, Roditi I 1994. Control of polyadenylation and alternative splicing of transcripts from adjacent genes in a procyclin expression site: a dual role for polypyrimidine tracts in trypanosomes? *Nucleic Acids Res 22*: 1359-1364.

Vellard M, Soret J, Viegas-Pequignot E, Galibert F, Nguyen VC, Dutrillaux B, Perbal B 1991. C-myb proto-oncogene: evidence for intermolecular recombination of coding sequences. *Oncogene 6*: 505-514.

Worthey EA, Martinez-Calvillo S, Schnaufer A, Aggarwal G, Cawthra J, Fazelinia G, Fong C, Fu G, Hassebrock M, Hixson G, Ivens AC, Kiser P, Marsolini F, Rickel E, Salavati R, Sisk E, Sunkin SM, Stuart KD, Myler PJ 2003. *Leishmania major* chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene. *Nucleic Acids Res 31*: 4201-4210.

Yang L, Lee O, Chen J, Chang CC, Zhou P, Wang ZZ, Ma HH, Sha HF, Feng JX, Wang Y, Yang XY, Wang L, Dong R, Ornvold K, Li BL, Chang TY 2004. Human acyl-coenzyme A:cholesterol acyltransferase 1 (acat1) sequences located in two different chromosomes (7 and 1) are required to produce a novel ACAT1 isoenzyme with additional sequence at the N terminus. *J Biol Chem 279*: 46253-46262.

Yu Y-T, Scharl EC, Smith CM, Steitz JA 1999. The growing world of small nuclear ribonucleoproteins. In RF Gesteland, TR Cech, JF Atkins (eds), *The RNA World*, Cold Spring Harbor Press, New York, p. 487-523.

Zeiner GM, Sturm NR, Campbell DA 2003. The *Leishmania tarentolae* spliced leader contains determinants for association with polysomes. *J Biol Chem 278*: 38269-38275.

Zhang C, Xie Y, Martignetti JA, Yeo TT, Massa SM, Longo FM 2003. A candidate chimeric mammalian mRNA transcript is derived from distinct chromosomes and is associated with nonconsensus splice junction motifs. *DNA Cell Biol 22*: 303-315.

Zorio DA, Cheng NN, Blumenthal T, Spieth J 1994. Operons as a common form of chromosomal organization in *C. elegans*. *Nature 372*: 270-272.