Trans-splicing of pre-mRNA is predicted to occur in a wide range of organisms including vertebrates

Thomas Dandekar and Peter R. Sibbald
European Molecular Biology Laboratory, Postfach 10 22 09, Meyerhofstrasse 1, D-6900 Heidelberg, FRG

Received June 8, 1990; Accepted July 5, 1990

ABSTRACT
Several known trans-splicing RNA structures were used to define a canonical trans-splicing structure which was then used to perform a computer search of the EMBL nucleotide database. In addition to most known trans-splicing structures, many putative new trans-splicing sites were detected. These were found in a broad range of organisms including the vertebrates. Control experiments indicate that the search predicts known false positives at a rate of only 20%. Trans-splicing may therefore be a very wide-spread phenomenon.

INTRODUCTION
When trans-splicing of pre-mRNA in trypanosomes was first reported [1] it was viewed as yet another peculiarity in an group of organisms already known to be atypical [2]. (For reviews of trans-splicing see [3-7].) The subsequent discovery of trans-splicing in chloroplasts [8-14] and nematodes [7] indicated that the phenomenon was much more widespread. Naturally the question arises; how widespread is trans-splicing? This question is enticing not only for academic reasons. As Boothroyd [15] has pointed out, many drugs that might control parasitic trypanosomes or nematodes also injure the host. If trans-splicing were to occur only in certain groups of organisms, (and particularly not in humans or cattle) then the trans-splicing reaction might provide an ideal target for novel drug therapies. The answer would of course also provide insight into the the more fundamental question of the evolution of trans and other forms of splicing [6].

While the 5' mini-exon that is trans-spliced in kinetoplastids is sufficiently conserved [16] that it was possible to biochemically locate Crithidia fasciculata mini-exons using a T. Brucei probe [17], there is sufficient divergence between the mini-exon sequences of nematodes and trypanosomes [18] to prevent the location of nematode mini-exons using the same methodology. In general for phylogenetically distant species (which are the interesting ones) it will not be possible to use a mini-exon probe to biochemically screen novel species as a way of discovering if they also trans-splice. Since it is possible to identify cis-splicing sites using computer searches [19], it seemed that such an approach might also be used to locate trans-splicing sites. Studies on the mechanism of trans-splicing have identified many important features of trans-splicing RNA structures [6, 20]. By using such features to search DNA sequence data bases, we have been able to detect new putative trans-splicing sites and present evidence that trans-splicing occurs in organisms not previously known to exhibit trans-splicing.

MATERIALS AND METHODS
The EMBL nucleotide sequence database 22.0 [21] consisting of 38 x 10^6 base pairs and 32 x 10^5 sequences was used. Searches of both strands were performed on the EMBL VAX cluster using Pascal programs custom-written by one of us (TD) for the purpose.

The target for which to search was derived from the six trans-splicing structures shown in Fig. 1. These structures were chosen because they are well documented [20] and relatively well understood. The canonical structure which was used as the target is shown schematically in Fig. 2. Features which were deemed obligatory were (1) the G-G doublet pairing with the Y-Y doublet; (2) a loop size of 3-10 nucleotides of which at least 3 must be U; (3) of the 4 positions following the G-G doublet, at least 3 of them must base pair with the opposite strand (here and elsewhere G:U is considered a pair). If 1,2 and 3 pair then the stem is extended only until 50% of the bases (including positions 1,2 and 3) pair. If one of 1,2 or 3 do not pair then the stem is extended until a non-pair is encountered. (4) The distance ranges, 0-7 bases are obligatory. (5) The Sm-site consists of a stretch of at least 3 U interrupted by 0 or 1 other nucleotides, bracketed at both ends by the doublet R-R. The first nucleotide of the Sm-site to be within 60 nucleotides of the trans-splicing loop. (6) Stem loops I and II were identified with a simple energy scoring scheme. Each G:C pair scored 3, A:U scored 2, G:U scored 1, and non-pairs scored -2. A one nucleotide bulge was permitted and scored -2. The resultant energy sum had to exceed both 6 and (the number of nucleotides in the stem plus loop divided by 2, rounded down to the preceding integer).

Non-obligatory target features also contributed to the evaluation of putative hits (a hit is a positive located by a search). Each of the following seven possible features contributed one point; (1) the first residue after the G-G is a U; (2) the second residue is an A; (3) on the the other strand of the stem the 5' most residue should be G or U; (4) the 3' adjacent residue is a U (5,6) the next two 3' adjacent residues are A or U; (7) there was no constraint on the next 3' adjacent residue, but the one after should be a U. At least 6 of the 7 non-obligatory features were required. Note that the stem labelled as 'non-obligatory features' may
overlap with positions labelled 1 to 4 (and its complementary strand) but the two have been drawn as non-overlapping in the interests of clarity.

RESULTS AND DISCUSSION

There were 327 hits with 6 of the obligatory features and 37 hits with all 7. The distribution of the hits in various groups of organisms is summarized in Table 1. Detailed information about each individual putative trans-splicing site is presented in Table 2. Trans-splicing sites are predicted both in introns and exons with a trend to have more intron examples in higher organisms. Most sites are either known to be transcribed as RNA (EX, IN, LT in table 2a and 2b; on (C) in Table 2b) or to exist as RNA as part of their life cycle (labeled as int in Table 2). It has been proposed that the DNA strand opposite a coding sequence also may often be transcribed [22] and many predicted trans-splicing sites are opposite CDS (Table 2).

Controls

Of the six sequences shown in Fig.1 (those used to build the canonical structure) the five stored in the EMBL database were correctly identified and had all 7 non- obligatory structures (L. enrietti was not stored in the database as an unsplit motif). In addition, several other known trans-splicing RNAs which were not part of the training set were correctly identified: LSMEDRNA (all idcodes refer to the EMBL database) and LSILINS1 from Leptomonas seymouri, TRSLRC and KTKPMC02 from T. cruzi; ALRIASl from Ascaris lumbricoides; CBRR5B, CBRRSA from Caenorhabditis briggsae; and CERR5 from Caenorhabditis elegans. In addition, numerous sites were identified in chloroplasts (Table 2) in which trans-splicing is known to occur [8-14]. All the known trans-splicing organisms including trypanosomes, Nematodes and Chloroplasts are correctly found by the search, usually with a score of 6 or 7. However, some trans-splicing sites were not found and these are discussed next.

Trans-splicing RNAs from Trypanosoma brucei which have a diverged Sm-site, RRTCRR [1] are not found (although KTKPCYB from T. brucei with a canonical Sm-site is found, Table 2). Use of the diverged T. brucei Sm-site makes the search quite non-specific (data not shown). Apart from this, trans-splice sites that are interrupted by introns, e.g. TCMXA or stored in the database as two or more parts such that the trans-splicing motif is split (e.g. L. enrietti, LESL1) are also not detected. Similarly, in some cases the GG of the trans-splicing structure is in the database but the up or down stream sequence does not appear in the sequence entry (e.g., CFMIEX, see also TCSLGB and TCSSLGA which stop at the first G). As would be expected, such truncated' sequences are not identified by the search. Finally there are some known trans-splicing RNAs for which the exact site is not known experimentally and the search should allocate a splice site for these. Rsps12 RNA in tobacco found by Keller et al. [12] to be trans-spliced, was missed. However a site was correctly predicted for the known trans-splicing of rps12 in liverwort; i.e., a site with 5 non-obligatory features is located 268 bases after rpl20 / rps12 exon1 on the complementary strand at base 65539 in a long transcript [23]. In Chlamydomonas reinhardtii psbB is known to trans-splice [13] but the search does not find a trans-splicing site. However the search does predict a trans-splicing site in psbC with 5 of the non-obligatory sites. Moreover, a trans-splicing site (with all 7 non-obligatory features) occurs in the Liverwort plastid at 42738, 14 base pairs upstream of psbB and another site occurs with 5 non-obligatory features in psbB (C) in tobacco at 39201. Four other sites in C. reinhardtii have also 5 of the non-obligatory features and thus are excluded from Table 2 (only hits with 6 or 7 of the non-obligatory sites are presented). A complete list of the search results including hits with 5 of the non-obligatory features is available on request from TD. If the hits with a score of five are included, the search identifies all genera known to trans-splice which are stored as a complete motif in the data base. From chloroplasts it is known that both complete exons and small leader sequences are trans-spliced. Thus we did not restrict our search by demanding an additional homology to the small spliced leader as this hampers the identification of known chloroplast sites (data not shown).

Table 1 indicates considerable variation in the number of predicted sites per 10^6 base pairs. From this it is tempting to conclude that trans-splicing is much more likely to occur in organelles, invertebrates and fungi than in the other groups. Caution is needed in making any such interpretation due to the presence of confounding factors. In particular, trypanosomes
Table 1. The occurrence of hits in various categories of DNA in EMBL 22.0. This database is divided, approximately along taxonomic lines, but primarily for convenience, into the categories: synthetic, viral, phage, organelles, prokaryotic, fungi, vertebrates, plants, vertebrates, mammals, rodents, primates, and unannotated. This Table shows the frequency of occurrence of hits with 6 or 7 of the non-obligatory positions (strong positives) in each category. The total amount of DNA in base pairs and the number of sequences in each category are also shown. The rightmost column is the number of hits (6+7) per 10^6 base pairs.

| Category          | sequences | base pairs | 6 point hits | 7 point hits | hits/10^6bp |
|-------------------|-----------|------------|--------------|--------------|-------------|
| synthetic         | 755       | 274 405    | 1            | 0            | 4           |
| viral             | 2 762     | 4 502 866  | 23           | 1            | 5           |
| phage             | 312       | 613 323    | 3            | 1            | 7           |
| organelle         | 1 550     | 2 191 964  | 52           | 16           | 31          |
| prokaryote        | 3 765     | 5 307 864  | 30           | 1            | 6           |
| fungus            | 1 455     | 2 143 702  | 31           | 0            | 14          |
| invertebrate      | 2 251     | 2 714 876  | 33           | 7            | 15          |
| plant             | 1 462     | 1 988 708  | 18           | 1            | 10          |
| mammal            | 1 172     | 1 468 528  | 6            | 0            | 4           |
| rodent            | 3 591     | 5 763 874  | 32           | 5            | 6           |
| primate           | 5 466     | 6 664 208  | 6            | 6            | 10          |
| unannotated       | 3 195     | 2 811 783  | 27           | 1            | 10          |
| total             | 31 308    | 38 234 565 | 327          | 37           | 10          |

Table 2a. Putative splicing sites from the strand provided in EMBL 22.0. Each category of DNA is headed by the name of the category followed by the percentage of the database which the category constitutes, followed by the percentage of the total hits, followed by the ratio. This provides a useful guide as to how frequently trans-splicing sites are relative to the amount of DNA being searched. After the heading the format is: EMBL idcode, position of GG in bases from the 5' end of the sequence (counting only a, u, c, g, t), sequence title (which may have been abbreviated), and comments. Between the idcode and the position an asterisk (*) indicates that all 7 of the non-obligatory features were found and that, as a consequence, this is a strong candidate for a trans-splicing structure. The absence of an asterisk means that 6 of the non-obligatory features occurred. If the same sequence is cited in several data base entries, their idcodes are given directly after the sequence title. The following conventions were used for the comments: rep = repetitive DNA; GG: means the splice site is found exactly at the 5' guanosine of the catalytic Guanosine diphosphate; test = indicates that the trans-splice site is one of the ones used to build the canonical structure; stt indicates a known trans-splicing RNA not from test set; knownO means that the organism is known to possess the trans-splicing reaction; knownS: known splice site; int: close to integration events; invA: before chloroplast inverted repeat A; pRNA: near tRNAs; cytb = in cytochrome B; Cox: in cytochrome oxidase subunit I; PCK: = protein kinase C; ori = origine of replication; nuc = nucleolus; km = kanamycin resistance; EX (ORF number): exon (open reading frame, number); IN (letter): intron(number); dn: after translation but before polyadenylation; pa: after polyadenylation site; ig: intergenic, no transcription unit. Fint,FinvA,PfinRNA,F: five prime of int, invA, start of mRNA transcription, respectively denotes a long RNA transcript occurring 3' from the trans-splicing site. For references to individual sequences and annotation, refer to the EMBL database (Cameron 1988).
The image contains a page of text, likely from a scientific journal, discussing various genes and their functions. The text is dense and includes multiple references to different species and gene names. The page is divided into sections, likely indicating different parts of the discussion, such as Vertebrates and Nucleotides. The text is fragmented and difficult to read due to the layout and formatting, but it appears to be discussing genetic sequences and their implications.

There is also a table labeled as Table b, which seems to outline trans-splicing sites from the strand complementary to that given in EMBL 22.0. This table is followed by a note stating that the position is where the GG occurs in strands given in EMBL 22.0, indicating a focus on specific DNA sequences and their implications.

The text and table are essential for understanding the genetic and molecular biology discussed in the page. The table provides specific data that can be analyzed for further research or educational purposes.
Phylogenetic distribution of predicted trans-splicing structures

Even taking into account that 20% of the putative hits in Table 2 might be false positives, there is still considerable evidence that trans-splicing occurs in several groups of organisms in which this mechanism has not been previously identified. There are groups in which it is not predicted such as the mycoplasmal but this may well be due to the small amount of mycoplasma DNA in the database. Despite the availability of $38 \times 10^6$ base pairs, this is actually a very small sample for this type of study. However, in general, the phylogenetic distribution of predicted trans-splicing structures is sufficiently broad to suggest that trans-splicing is quite primitive. The question has been raised as to whether trans-splicing is an unusual type of splicing that evolved in trypanosomes and a few other organism as an adaptive feature [6]. We would argue that trypanosomes did not develop trans-splicing as an adaptation but that they have retained it.

We observe that in a high number of cases (40), the predicted trans-splicing site is proximal to an integration site and that in an additional 12 cases, the predicted trans-splicing site is in or proximal to repetitive elements or transposons. Other workers [25,26] have observed an association between retroposons and mini-exons. This association of trans-splicing structures and integrating DNA may synergistically accelerate the spread of both but perhaps also contributes to the recombination of protein coding regions originally carried by the RNA having the respective trans-splicing site.

Strong candidate sites

It seems appropriate to identify some particularly strong candidates for experimental testing. In particular, MIDMURFV, SMTRNA1, LHDEL, PCTHYSY, GGMYHE, HSKCB1A, and HSTCVA5 contain the catalytic double guanosine [20] exactly at an exon boundary. The ox13 locus in yeast (MISC13) has a well formed predicted trans-splicing site in the intron a4. The intron is already known to be important for splicing [27] and the following self-splicing group II intron a5g could be divided in vitro to yield two RNAs that trans-spliced in vitro with associated trans-branching of excised intron fragments [28]. Refer to Fig. 3 for secondary structure diagrams of three of these strong candidate sites. Comparison of these with the structures in those in Fig. 1 shows how highly similar in structure they are to known sites. More examples are given in Table 2; particularly striking are cases in which a trans-splicing site is found in similar positions in the same gene from different organisms (class in Table 2) or additional evidence is available (Table 2 and notes to Table 2).

CONCLUSION

The search identified the five trans-splicing structures from the test-set which are undisrupted in the EMBL database and detected (with correctly predicted trans-splicing sites) all groups known to trans-splice, including Trypanosomes, Nematodes and Chloroplasts. Neither rRNAs (two exceptions), tRNAs nor small nuclear RNAs involved in cis-splicing were mistakenly identified as trans-splicing sites. The search could not identify every known trans-splicing site from every species (deregarded T. brucei sites, two known Chloroplast trans-splicing RNAs and truncated trans-splicing sites in data base entries were missed). Other RNA structures also could promote trans-splicing and might not have been detected. It is also possible that some putative sites are in fact pseudogenetic in nature. However, enough new

Figure 3. Three strong candidates for newly identified trans-splicing sites. The secondary structures of these predicted sites are very similar the known sites shown in Fig.1 HSKCB1 is human protein kinase C, MISC13 is yeast mitochondrial, and GGMYHE is chicken embryonic myosin heavy chain. None of these organisms are known yet to trans-splice in vivo.

appear to trans-splice every pre-mRNA [3,6,16,24] and so the observation that there are many more splice sites per unit sequence in trypanosomes may be simply because they splice a higher fraction of their pre-mRNA than do other organisms. Secondly, there is a tendency in molecular biology to sequence DNA related to previously sequenced DNA and whether this is due to the availability of particular probes or common interest in certain sequences, the result is that the database does not consist of ‘independent events’.

The search was able to distinguish between trans-splicing sites and other types of RNA (Table 2). In no cases did the search mis-identify tRNA as a trans-splicing site (there is however a plausible site which would release tRNA (proline) from a long precursor in Spiroplasma, SMTRNA1). Only in two cases (N. gruberi 18S RNA, and L. tarentolae mitochondrial 125 rRNA) were rRNA genes apparently mistaken for trans-splicing sites and no small nuclear RNA known to participate in cis-splicing was confused with a trans-splicing site. Both cis- and trans-splicing RNAs have several similar features, but perform different biological functions and the search discriminates between them. These three negative controls for RNA structures which appear similar but which are functionally distinct underline the ability of the search to pick good candidate sites for trans-splicing.

In an attempt to estimate the background error rate of the search, the search was repeated on the strand in the EMBL database with the sequence YY(≥ 3 C, ± 1 non C)YY substituted for the Sm-site. This motif should be biological nonfunctional as its Sm-site is destroyed [20]. This search found 12% as many ‘hits’ with 6 or 7 of the nonobligatory features as did the search using the correct Sm-site. We estimate therefore that as many as 80% of the putative hits in Table 2 may be real. Similarly, since it is known from T. brucei that other trans-splicing sites can exist with diverged Sm-sites, and possibly with other variations of which we are as yet unaware, the search is almost certainly incomplete.
candidate trans-splicing sites (even taking into account a background estimate of 20% false positives) have been detected to suggest that trans-splicing may be much more widespread than previously thought. There are several good candidate structures identified in species not yet known to possess trans-splicing available for experimental testing (Table 2), including sites from vertebrates.

ACKNOWLEDGMENTS

We thank Angus Lamond, David Tollervey and Benjamin Blencowe for reading the manuscript and making suggestions. PRS is grateful to the National Sciences Engineering and Research Council of Canada and the Alexander von Humboldt-Stiftung for financial support. TD wishes to thank Boehringer Ingelheim Funds for Basic Medical Research for support.

REFERENCES

1. Milhausen, M., Nelson, R.G., Sather, S., Selkirk, M. and Agabian, N. (1984) Cell 38, 721–729.
2. Sharp, P.A. (1987) Cell 50, 147–148.
3. Borst, P. (1986) Ann. Rev. Biochem. 55, 701–732.
4. Van der Ploeg, L.H.T. (1986) Cell 47, 479–480.
5. Braun, R. (1986) Bioessays 5, 223–227.
6. Laird, P.W. (1989) Trends. Genet. 5, 204–208.
7. Nilsen, T.W. (1989) Exp. Parasitol. 69, 413–416.
8. Ohyama, K., Fukazawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H. and Ozeki, H. (1986) Nature 327, 572–574.
9. Shinozaki, K., Ohme, M., Tanaka, M., Watanuki, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chungwongse, J., Obakata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B.Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusada, J., Takaawa, F., Kato, A., Tohdoi, N., Shimada, H. and Sugiyama, M. (1986) EMBO J. 5, 2043–2049.
10. Umesono, K., Ozeki, H. (1987) TIG 3, 281–287.
11. Zaita, N., Torazawa, K., Shinozaki, K. and Sugiyama, M. (1987) FEBS Lett. 210, 153–156.
12. Koller, B., Fromm, H., Galun, E. and Edelman, M. (1987) Cell 48, 111–119.
13. Kuck, U., Choquet, Y., Schneider, M., Dron, M. and Bennoun, P. (1987) EMBO J 6, 2185–2195.
14. Hiratsuka, J., Shinada, H., Whittier, R., Ishihashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, C.R., Meng, B.Y., Li, Y.Q., Kanno, A., Nishizawa, Y., Hirai, A., Shinozaki, K. and Sugiyama, M. (1989) Mol. Genet. 217, 185–194.
15. Boothroyd, J.C. (1985) Ann. Rev. Microbiol. 39, 475–502.
16. De Lange, T., Berkvens, T.M., Veerman, H.J.G., Carlos, A., Frasch, C., Barry, J.D. and Borst, P. (1984a) Nucl. Acids Res. 12, 4431–4443.
17. Muhich, M.L., Hughes, D.E., Simpson, A.M. and Simpson, L. (1987) Nucl. Acids Res. 15, 3141–3153.
18. Krause, M. and Hirsh, D. (1987) Cell 49, 753–761.
19. Senapathy, P., Shapiro, M.B. and Harris, N.L. (1990) Meth. Enzymol. 183, 252–278.
20. Bruzik, J.P., Van Doren, K., Hirsh, D. and Steitz, J.A. (1988) Nature 335, 559–562.
21. Cameron, G.N. (1988) Nucl. Acids Res. 16, 1865–1867.
22. Tramurano, A., Scalfato, V., Barni, N., Cipollaro, M., Franze, A., Macchiato, M.F. and Cancino, A. (1984) Nucl. Acids Res. 12, 5049–5059.
23. Kohchi, T., Obara, Y., Umesono, K., Yamada, Y., Komano, T., Ozeki, H. and Ohyama, K. (1988) Curr. Genet. 14, 147–154.
24. De Lange, T., Michels, P.A.M., Veerman, H.J.G., Cornelissen, A.W.C.A. and Borst, P. (1984b) Nucl. Acids Res. 12, 3777–3789.
25. Affoter, M., Rindlsbacher, L. and Braun, R. (1989) Gene 80, 177–183.
26. Aksoy, S., Lalor, T.M., Martin, J., Van der Ploeg, L.H.T. and Richards, F.F. (1987) EMBO J. 6, 3819–3826.
27. Dujardin, G., Jacq, C. and Slonimski, P.P. (1982) Nature 298, 629–632.
28. Jarrell, K.A., Dietrich, R.C. and Perlman, P.S. (1988) Mol. Cell Biol. 8, 2361–2366.
29. Miller, S.I., Landfear, S.M. and Wirth, D.F. (1986) Nucl. Acids Res. 14, 7341–7360.