© 1990 Oxford University Press

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 Boothyroyd [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 fascuculata* 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 38×10^6 base pairs and 32×10^3 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 until only 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 had 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

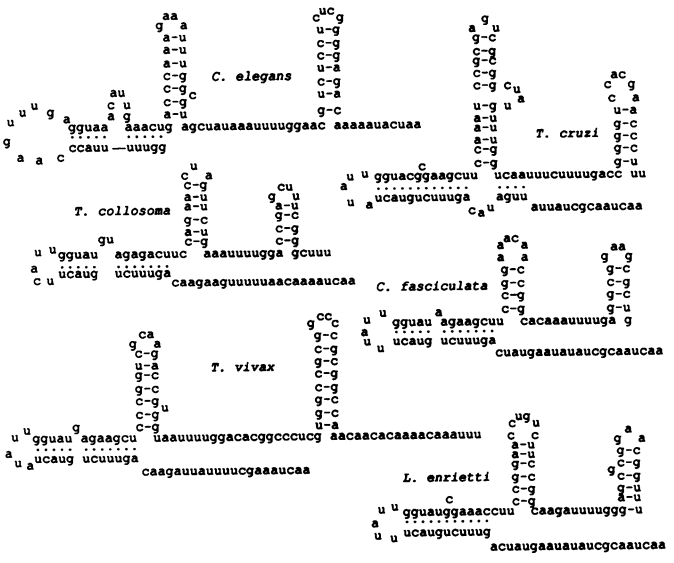

Figure 1. The 6 *trans*-splicing structures used to build the canonical structure, redrawn from [20]. These structures have been described in detail as follows, *Caenorhabiditis elegans* [18], *Trypanosoma cruzi* [16], *Crithidia fasciculata* [17], *Leptomonas collosoma* [1], *Trypanosoma vivax* [16], *Leptomonas enrietti* [29].

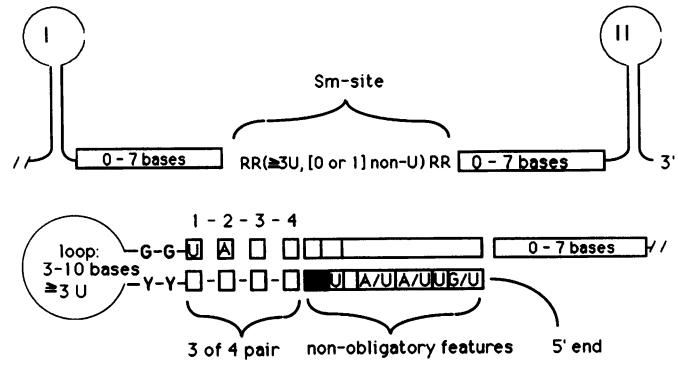

Figure 2. The canonical *trans*-splicing structure that was used as a target to perform searches. See Materials and Methods for details.

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*; ALRLASL from *Ascaris lumbricoides*; CBRR5B, CBRR5A 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, RRTCTRR [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 TCSLGA 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. Rps12 RNA in tobacco found by Koller 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 reinhardii* 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. reinhardii* 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 occurance 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, invertebrates, plants, vertebrates, mammals, rodents, primates, and unannotated. This Table shows the frequency of occurance 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's'up3(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	512	613 323	3	1	7
organelle	1 550	2 191 964	52	16	31
prokaryote	3 765	5 301 864	30	1	6
fungi	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
vertebrate	1 563	1 794 464	14	0	8
mammal	1 172	1 468 528	6	0	4
rodent	5 591	5 763 874	32	3	6
primate	5 466	6 664 208	57	6	10
unannotated	3 195	2 811 783	27	1	10
total	31 508	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 course estimate of how frequent *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 nonobligatory 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 nonobligatory 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 catalytical Guanosine doublette; test indicates that the *trans*-splice site is one of the ones used to build the canonical structure; ntst 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; ptRNA: near tRNAs; cytB = in cytochrome B; cox: in cytochrome oxidase subunit 1; PKC = protein kinase C; ori = origine of replication; nuc = nucleoline; km = kanamycine 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,FmRNA,F: five prime of int, invA, start of mRNA transcription; LT 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).

Synthetic

ECRGNABP	604	E.coli rrn promotor/terminator fusion		ig		fusion

Viruses

CORIBASP	5450	Avian infectious bronchitis virus		mid	EX	LT
EBJNC1	244	Epstein-Barr integration;	int	360T	Fint	
HSSLJT1	244	Epstein-Barr integration;	int	360T	Fint	
HANSNC	331	Hantaan virus nucleocapsid protein		mid	EX	LT
HEHSLTA	119	Herpes ateles thymidylate synthase	int	142T	FmRNA	
HEHSSTS	469 *	Herpes saimiri thymidylate synthase	int	102T	FmRNA	
HEVZVXX	68271	Varicella-Zoster virus(95724 272T up)		325T	EX	LT
NCBNYVV1	3408	Beet Necrotic Yellow Vein Virus RNA-1		mid	EX	LT
PA16	6391	Human papillomavirus type 16 (HPV16)		mid	EX	LT
PAPPPH31	4103	Human papillomavirus type 31 (HPV-31)		mid	EX	LT

Phages

INPF3C01	1884 *	Bacteriophage Pf3; (NewYork strain)		80F	EX ORF	93
INPF3COM	1884	" (Nijuegen strain)		80F	EX ORF	93
POP22INT	373	Lambdoid phage P22 int-xis region		ig		

Organelles

ALCRDNA	3592	Astasia longa chloroplast ribos.DNA		ig,not in rRNA		
CFRPME	416 *	C.fasciculata mini-exon repeat	test	GG		control
CHEGS16R	2277	Euglena grac. chloroplast rRNA dupl.	int	0T	int 2	
CHHVPSBD	3221	Barley chloroplast psbB,H;petB,D	cytB	20T	up	LT,600
CHMPXX	69971 *	Liverwort Marchantia chloroplast	psbB	581T	EX 1	knownO
CHMPXX	71381	(12 Trans5)	cytB	43T	up	knownO
CHMPXX	109520	(99837 ig)	invA	140T	FinvA	knownO
CHNTXX	60343	Tobacco chloroplast; (5 Trans5)	558F	EX ORF	512	knownO
CHNTXX	130104	(152057 ptRNA,ig)	invA	400T	FinvA	knownO
CHOSXX	99488	Rice complete chloroplast genome		400F	dn ORF	23
CHSARS16	166	mustard chloroplast rps16 gene		290T	up	LT
CHTATRN1	316	Wheat chloroplast,URF62,5tRNA genes	ptRNA,EX		URF62,LT,note1	
CHZMRN4	28	Maize Arg- and Asn-tRNA 3' region	ptRNA		ig	
KTKPCYB	508	T.brucei kplast apocyt.b	EX,note2,knownO			
MITB01	1100	"	EX,note2,knownO			
KTKPMC02	609	T.cruzi kplast minicircle DNA pTc-21	ntst		GG	knownS
LSMEDRNA	460	Leptom. seymouri mini-exon	ntst		GG	knownS
LSLINS1;	782	" mini exon with insert element LINS1	ntst		GG	knownS
MIBSRNAL	118	Boletus satanas mt large rRNA gene		EX		
MICAARS	1272	Cephalosporium acremonium mtDNA ARS	ori		ig	
MIDMURFV	68	D.virilis mitochondrial DNA rRNA/URF	GG exactly at boundary			
MIDYTRN	1173	D. yakuba mitochondrial DNA	ori		ig	
MILTRRNG	525	Leishmania tarentolae mtDNA(12S rRNA)		90F	EX	
MINCND	260	Neurospora crassa mt NADH dehydr.ase	NADH	160T	up	LT, note7
MIPRGPL	3053	P. primaurelia mt rDNA; 3053; 3053;	ori		ig	
MISC13	7855	yeast cytochrome oxidase subunit1	cox1 mid		IN aI4	text
MISC23;	522	as above				
MISCCO1;	522	as above				
MISPXX	5991	Sea urchin mt genome (Trans5 in NADH)	cox1	210F	EX	
MITGTRN1	196	Torulopsis glabrata mt tRNA genes	ptRNA		ig	
SCMTOR7A	51	Yeast (petite) mt replication origin	ori		ig	
TCMIEX1	35 *	Trypanosoma cruzi mini-exon repeat	test		GG	control
TRSLRC	80 *	T. cruzi small spliced leader	ntst		GG	knownS
TRSLRLC	77 *	T.Leptom. coll. small splic. leader	test		GG	control

Prokaryotes

ASCHMO	61	Acinetobacter cyclohexanone monooxyg.		327T	FmRNA	
ATACH5	18611	A.tumefaciens plasmid pTi15955 T-DNA		54T	EX ORF23	note4
BSPRBH1K	1510	Therm. bacillus plasmid pRBH1 (km)	km	123F	EX	LT
CTORF	2231	C.trachomatis plasmid pCTT1; CTDNAB;		mid	EX ORF2	
ECBIRA	60	E.coli biotin birA gene		238T	FmRNA	
ECCPELC	478	Erwinia carotovora endo-pectate lyase		31T	FmRNA	
ECDMS	6146	E.coli anaerobic dim. sulfoxide red.			ig	
ECRRNBZ	6674	E. coli 16S rRNA, tRNA and two urfs			ig	
HIOMPP6	535	H.influenzae outer membrane protein P6			ig	
MVMCR1	2932	Methanoc. voltae methyl-CoM reductase		228F	EX	
PMMB66EH	295	Plasmid pMMB66EH expression vector			ig	
PSIRM	498	P.stuartii pstI genes			ig	
SA110KAR	1510	S.aureus plasmid pUB110dB (km)	km	123F	EX	LT
SAPS194	797	S.aureus plasmid pS194			ig	

SMTRNA1	389	Spiroplasma meliferum tRNAs; 389;		GG at 5'tRNA		note3
SPLYTPN	131	S.pneumoniae autolysin gene		67T FmRNA		
VHCHIT	3179	V.harveyi N,N'-diacetylchitobiase; ig,16 bp 3' of 29 bp hairpin				
Fungi						
CAERG16	322	Candida albicans cytochrome p-450 L1A1		187F EX		
KLGAL	3997	Kluyveromyces lac. GAL1, GAL7 and GAL10		269F EX	Gal7	
KLK1P	3830	Kluyv.lactis killer plasmid k1		400F EX	ORF1	
KLK1P	6365	(killer toxin is ORF1); KLKILL05,KLKILL1L;		150T EX	ORF1	
SCDEL1	185	Yeast delta and truncated delta element	int	mid	int	
SCDP8	908	Yeast delta-P8 gene 5' region		mid	EX	
SCHAP2	43	Yeast transcriptional activator HAP2		97T	up	
SCHOMMT	121	Yeast nuclear dna homologous to mt dna			ig	
SCRPS31	935	Yeast gene for ribosomal protein S31		120T EX		note 10
SPMEI2	3318	S.pombe mei2 gene		604F pa		LT
SPTUBA1	764	S.pombe alpha-tubulin 1		383F EX		
Plants						
GMGY3	2466	Soybean glycinin subunit G3 gene		140F IN 3		
GMTGM1	2752	Soybean lectin transposon Tgm1	int	800T int		PI
LECHSOD	134	Tomato superoxide dismutase mRNA		102F EX		
LHDEL	6885	Lilium del transposon (6531 EX)	GG exactly at the STOP of exon			
MCPPCB	1801	M. crystallinum phosphoenolpyrcarb.		mid	IN 3	
PSLECPGA	225	Pea PSL2 lectin pseudogene	int	mid	int	
STPATP1	4900	Potato patatin pseudogene (SB6B)	int	mid	int	
STPATP2	5964	Potato patatin pseudogene (SA10C)	int	mid	int	
STWIN12G	3431	Potato wound-induced genes WIN2		300T mRNA		
ZMCPPS2G	25	Maize chloroplast psbG gene		130T up		
ZMZEI19	3206	Maize gene for Mr 19000 alpha zein		54F pa		
Invertebrates						
ALRGASL	544 *	A.lumbricoides spliced leader	ntst	GG	knownS	
BMCH01	861	silkmoth chorion protein Hc-B.13		mid	EX	
BMCHRHCA	1745	silkmoth chorion protein Hc-A.12/B.12		189F EX 2		
CBRR5B	336 *	Caenorhabditis briggsae 5S rRNA	ntst	GG	knownS	
CEACTL	352 *	C.elegans actin spliced leader	test	GG	control	
DHMIF8A	3877	Drosophila hydei microcopia dhMiF8		661F pa	/ ig	
DMGPDHA	3902	Drosophila GP-dehydrogenase; DMGPDHG		68T IN		
DMWHITE	13867	Drosophila white locus			ig	
NGRGE	799	N.gruberi 18S subunit rRNA gene		802F EX		
PCTHYSY	1267	P.carinii thymidylater synthase gene	GG is exactly at EX4/IN D			
PFANT2L	409	P.falciparum antigenic determinant		30T up		note9
PFRSI	1361	P.falciparum repetitive DNA	int	mid	int	
PFTRAP	266	P.falciparum thrombospondin rel.prot.		30T up		note9
TTCNJB	2114	Tetrahymena thermophila cnjB gene		200F IN 5		
TVMIEX1	35 *	Trypanosome vivax mini-exon repeat	test	GG	control	
Vertebrates						
CIACTB	3965	Grass carp beta-actin gene		643F dn		
GGEF9E3	898	Chicken embryo fibroblast protein mRNA		284T dn		LT
GGMYC	3384	Chicken cellular myc onc.gene;GGCMYCA		mid	IN 2	
GGMYHE	8908	Chicken embryonic myosin heavy chain	GG exactly at IN13/EX14			
GGPGR	4270	Chicken progesterone receptor mRNA		30T dn		
GGRSVIND	878	Chicken RSV-transformed mRNA		400T dn		
XLENK02	40	Xenopus laevis proenkephalin gene A2		56T IN 1		
XLLAMLI	2366	Xenopus mRNA for nuclear lamin L(I)		100T dn		
Mammals						
CHEBGLII	1019	Goat epsilon II beta-globin gene		90F IN 2		
OCPKCBR	2237	Rabbit protein kinase C beta mRNA	PKC	7F dn		
Rodents						
DOREPI	2778	Kangaroo rat repetitive DNA; 2779;	int	516T int		LT
MAHPRT	1083	Chinese hamster hprt mrna		60T dn		
MMBGLOFG	45	Mouse downstream of beta-globin gene		600F pa		LT
MMDHF4	781	Mouse mutant tetrahydrofol.-DH mRNA		150T dn		
MMDHF7	325	Mouse dihydrofolate reductase exon3		10T IN 2,near EX		
MMETNB	2272	Mouse early transposon (ETn)	int	30T Fint		
MMIRF12	2192	Mouse interferon reg. factor-2 mRNA		243T dn		LT
MMNUCL01	1906	Mouse nucleolin gene ; MMNUCLEO;	nuc	290T IN 1		
MMPP105R	3880	Mouse retinoblastoma susceptib.; 3885;		1020F dn		
MMT1CPS	2003 *	Mouse T1c pseudogene for MHCI antigen		239F IN 3		
RNCAMI3	1915	Rat CaMI gene for calmodulin		mid	IN 3	
RNFBAG	3420	Rat gene for alpha-fibrinogen		573T IN 1		
RNNUCLEO	2064	Rat nucleolin gene, exons 1 and 2	nuc	290T IN 1		
RNPECOA	1509	Rat peroxisomal enoyl-CoA mRNA		578T EX		
RSLIN4A	3033	Rat long interspersed repet. DNA	int	mid	int	
Primates (all hits are human)						
HSB2M2	2113	Human beta-2-microglobulin gene		103T IN 3		
HSC1A1	2428	Human alpha 1 collagen type I gene		56F IN E		
HSCRPG	1930	Human C-reactive protein; HSCRPGA;		500T dn		
HSEB2CR2	3943	EBV receptor cr2 RNA; HSEBUR14;HSEBVR ;		100T dn		
HSFIBEDA	2718	fibronectin gene ED-A region; 2719		278T IN 2		note6
HSGA7331	1691	pancreas CA marker mRNA GA733-1;HSGA733A;		102T dn		
HSGASTB	479	Human gastrin gene, 3' region		400F pa		LT
HSHBEG	8493	Human L1Heg repetitive element	int	mid	int	
HSHBVINT	314	Human DNA / hepatitis B virus integr.	int	600T Fint		
HSHK2A	3726 *	Human calcium-ATPase mRNA;HSCAATP4;		55T dn		LT
HSHMG14	650	non-histone protein HMG-14 mRNA		190F dn		LT
HSHMGCOB	1140 *	HMG CoA reductase (EX1 and promoter)		mid	IN 1	
HSHSP90B	3862 *	Human 90 kD heat shock protein		23F IN D		
HSHTV2A	1028	Human tRNA-Val family	ptRNA	100T FtRNA		
HSIFNB3	11512	Human interferon-beta-3 locus			ig	
HSIGVKA2	1910	Ig-kappa V(k)III pseudogene A22		150T IN 1		
HSINSRC	5420	Human insulin receptor allele 1; 5420;		724T IN O		Note5
HSINSRD	5019	Human insulin receptor allele 2; 5019;		724T IN O		Note5
HSIRF2	1947	interferon regulatory factor-2 mRNA		197T dn		
HSMHDRW1	2246	MHC class II HLA-DRw53-beta		1200F IN 1		
HSPKCB1A	2152	protein kinase C beta I mRNA; HSPKCB1;	PKC	GG at STOP Codon		
HSPROL1	478	Human prolactin gene 5' region		400T up		
HSRBS	3976	Human retinoblastoma susceptib. mRNA		600F dn		note8
HSRSKP08	311	Human kpni repeat mrna	int	mid	int	
HSTCGVA5	1044	T-cell receptor pseudogene; HSTCGVA;	GG at known splice site			
HSU6RNA	152	Human gene for U 6 RNA	int	69T upstream of U6		

note1: Long mRNA, transcribed together with preceding psbC and psbD. Cleavage at the GG would yield a thylacoid membrane spanning peptide.

note2: Long mRNA; apocyt.b RNA is split and spliced in aspergillus and yeast; The URFs surrounding T.brucei apocyt. b lack polyadenylation signals, but there is an A-rich sequence at the 3'end of a cDNA found by URF2 probe.

note3: The 10 bp upstream of tRNA Pro together with the tRNA Pro form a TRANS6 which exactly releases 5'end tRNA Pro from the long five tRNA precursor.

note4: The TRANS starts at bp 18616 and ends exactly at the end of ORF(bp18687).

note5: The two alleles have slightly different TRANS, but at the same position.

note6: EX2 before IN2 is untranslated in liver and alternatively spliced mRNAs with and without EX2 occur in different ratios in different tissues.

note7: Longer RNA species of 5, 5.6 and 7 kb are only lit by probes 5' to the nd4L gene or exon probes but not by intron or downstream probes.

note8: A trans-splicing event would lead to the shortened 4 kb mRNA observed in retinoblastomes.

note9: Identical positions. Parasite under genetic pressure like Trypanosomes!

note10: Entire 3'end (128 bp) of protein S31 mRNA forms a trans-splicing RNA.

Table b. *Trans*-splicing sites from the strand complementary to that given in EMBL 22.0. The position is where the GG occurs in terms of the strand given in EMBL 22.0; it is *not* the distance from the 5′ end of the complementary strand. Legend, as in Table 2a with the addition that CDS is coding sequence and opp means opposite from the strand searched. On (C) means that the transcription unit is here, on the complementary strand.

Synthetic(no hits)						
Viruses						
ADLE2B	1944	Adenovirus type 12 E2b region DNA	pol	mid	EX	on(C)
IRIEPEH	1190	Insect iridescent virus type 6	rep			
PXVACLEM	805	Vaccinia virus transposition mutant	int			
REMULVT3	447	Murine leukaemia virus (MuLV) 3'LTR	retrovirus			
REMULVT5	432	Murine Leukemia virus (MuLV) 5'LTR	retrovirus			
RESPUENV	377	Human spumaretrovirus	retrovirus			
COTGEV3	2941	Enteric coronavirus	opp non-structr. protein			
CORTGEVM	1020	Enteric coronavirus	opp nucleocapsid protein			
HRVVP2	864	Human rotavirus	opp vp2			
PARPVC01	3471	Canine parvovirus	opp vp2			
PARPVCCP	419	Canine parvovirus	opp CDS			
PARPVCVP	1799	Canine parvovirus				
MCACGDH	64	Cauliflower mosaic virus				
VSVNJNPA	195	Vesicular stomatitis virus N protein	opp CDS			
Phages						
MYOVP1	542	Bacteriophage P1 IS2 insertion hot spot	int	21F	ORF1	on(C)
STSPO2	486	Bacteriophage SP01 with terminal repeat	GG start EX		gp28	on(C)

Organelles

CHCERR23	759	Chlorella ellipsoidea plastid	int Tn like sequence
CHMPTRN	2960 *	Liverwort Marchantia plastid;	known0 mid EX ORF 704 on(C)
CHMPXX	46949	Liverwort Marchantia plastid;	known0 250F EX psaA on(C)
CHMPXX	30617		known0
CHMPXX	42738 *	very good candidate!!	known0 14F EX psaB on(C)
CHOSXX	115630	Rice complete chloroplast genome	between ORF63 and ORF23 on(C)
CHOSXX	103186	Rice complete chloroplast genome	NADH 451F EX ND5 on(C)
CHOSXX	49441	Rice complete chloroplast genome	NADH 221F EX ND3 on(C)
CHZMNDHD	263	Maize chloroplast ndhD, ndhE and psaC	NADH
MIDMM2	124 *	D.melanogaster mt large rrna gene;	GG at 17T EX on(C)
MIDMTRN	7367 *	Drosophila mt DNA;	NADH 24F EX ND1 on(C)
MIDYRRN	15258	Drosophila yakuba mt DNA	ori
MIMM01	1475	Mouse mtDNA	259F opposite 16S rRNA
MINCND2D	801	Neurospora crassa mt DNA duplications	int 259F NADH duplicat
MIPAIVS2	961	Podospora anserina class II intron	int homologous to RT
MIRC12S	1869	Rana catesbeiana mtDNA	988T opposite 16S rRNA
MIRNRN	280	Rat mtDNA; MIRNXX	255F opposite 16S rRNA
MISC13	2283	Yeast cytochrome oxidase subunit 1	cox1
MISCCO12	2656	Yeast cytochrome oxi1 gene and flanks	cox1
MISCORIK	281	Yeast mitochondrial ori2-ori7 region	ori
MITBCOX	1903 *	Trypanosoma brucei mt cyt c oxidase	82F EX on(C)
MITOMM	1351	Mouse mitochondrial genome	257F opposite 16S rRNA
MITOMM	10410	Mouse mitochondrion	opp URF4
MIXLG	3343	Xenopus laevis mt genome ; XLMTDTG	257F opposite 16S rRNA
MIXLORI	866	X.laevis mt ori	ori
CHNTXX	90472	Tobacco plastid	known0
MIBTXX	4916	Bovine mitochondrion	
TBGP01	571	T.brucei surface protein	

Prokaryotes

BAAPR	322 *	B. amyloliquifaciens alk. protease	
BSPRBH1K	198	Thermophile kanamycin plasmid	between reps
BSREPB	198	Bacillus plasmid	
BSRODC	3662	Bacillus subtilis rodC operon	opp CDS
ECCE1Z	75	Erwinia chrysanthemi endoglucanase	
MVMCR	4360	Methanococcus van.	
MVRPOP	6666	Methanococcus van.	
NGTIA	72	N.gonorrhoeae transformation inhibitory DNA	opp rpL15
SAL54BOP	220	S.aureus phage L54 attL site	
SAPUB110	2843	S.aureus plasmid	in neo(r) CDS
SMPAC	4797	Strepto.mutans	opp CDS
SMRPLKA	214	Serratia marcesc.ribos protein L11,L1	

Fungi

DDAAC11	975	Dictyostelium discoiedeum AAC-rich mRNA	opp CDS
DDACTA32	731	Dictyostelium discoiedeum actin	opp CDS
SCADE3	4176	Saccharomyces cerevisiae C-1-tetrahydrofolate synthase;	opp CDS
SCBAF1	1426	Saccharomyces cerevisiae transcription factor Baf1;	opp CDS
SCCPA1	1535	Saccharomyces cerevisiae carbamoyl-phosphate synthetase;	opp CDS
SCGCD1	1438	Saccharomyces cerevisiae GCD1 gene;	opp CDS
SCMAL28C	1961	Saccharomyces cerevisiae mutant mal2-8cp gene	opp CDS
SCMAL6R	1429	Saccharomyces cerevisiae MAL6R gene;	opp CDS
SCMAT4	294	Saccharomyces cerevisiae mating type;	
SCMYO1	592	Saccharomyces cerevisiae myosin-like cdc protein;	opp CDS
SCPDC1	1401	Saccharomyces cerevisiae pyruvate decarboxylase;	opp CDS
SCRAD50	3032	Saccharomyces cerevisiae RAD50 gene;	in heptad repeat region
SCRAR1	1869	Saccharomyces cerevisiae RAR1 gene;	
SCSERS	1432	Saccharomyces cerevisiae seryl-tRNA synthetase;	opp gene
SCSILA	244	Saccharomyces cerevisiae silencer DNA;	
SCSIR2	255	Saccharomyces cerevisiae mating type control);	

Plants

ASPHT3A	231	Avena sativa phytochrome	
GMLEA	843	Soybean lectin	transposon like sequence
HVLEU	911	Barley thiol protease	near many reps
IBGSPOA1	433	Sweet potato sporamin A	
PSELIP encoded	1702	Pea plastid early-light-induced protein	chloroplast nuclear
SCNACT	2112	S.cerevisiae N-acetyltransferase	opp CDS
VFVICG	2885 *	Vicia faba vicilin gene	

Invertebrates

CBRR5A	586 *	Caenorhabditis briggsae 5S rRNA (1kb)	known0 in spliced leader!
CERR5	210 *	Caenorhabditis elegans DNA for 5S rRNA	known0 in spliced leader!
CETUBUB	1631	Caenorhabditis elegans beta-tubulin	known0
DMANTPE8	497	D.melanogaster antennapedia; DMANTPRA	
DMIS176	6632	D.melanogaster copia-like element 17.6	int
DMIS297	6286	D.melanogaster transposable element 297	int
DMLGL2	5099	D.melanogaster giant larvae;	int these two sites
DMLGL2	2293	D.melanogaster giant larvae;	int are identical
DMRT412G	2740	Drosophila retrotransposon 412; gg at 2754 is also a site	
DNTN10P	2665	D.nebulosa transposon N10	int
PFIRAA	229	P.falciparum interspersed repeat antigen	int
DML2AMD	943	Drosophila α-methyldopa hypersensitivity	opp intron
DMESPLM7	217 *	enhancer of split	opp noncoding transcript
DMUBXG5	2679	ultrabithorax promoter	
PCMSA	755	P.chabaudi merozoite antigen	opp surface antigen CDS
PFSA27	2614	Plasmodium falciparum S-antigen	upstream of poly rep region
PYCSP	281	P.yoelii circum-sporozoite	opp CDS
PYCSP1	281	P.yoelii circum-sporozoite	opp CDS
SPC4X	831	Strongylocentrotus.purpuratus collagen IV	opp intr
TTHI01	1978	Tetrahymena H4-I gene and flanks	

Vertebrates

GGERBBF	2299	Chicken c-Erb oncogenic ALV insertion;	int
GGC1A225	525	Chicken alpha-2 collagen I	
GGOVAL	6224	Chicken ovalbumin gene;	opp intron G
GGPEC	310	Chicken ppenolpyr. carboxykinase	

Mammals

BTNABGSA	1307	Bovine galactosyltransf.	upstream of coding region
OCIL1R	1958	Rabbit interleukin 1 pre-cursor	
OCPRG5	734	Rabbit progesterone receptor	
SSAPOB2	3621	Pig apolipoprotein B	opp exon

Rodents

MMCY01	874	Mouse cytochrome P3-450;	
MMCYP245	837	Mouse cytochrome P3-450;	opp CDS
MMLVPA	2452	Mouse endogenous retrovirus;	int, all are nearly identical sequences
MMERMB56	391	"	
MMERMB73	390	"	
MMMLVPA	2452	"	
MMERU3L6	396	"	
MMU3LTRB	464	" rep in LTR region	
MMLTRIS	427	" has inserted CTR-IS element	int, rep
RNRL13	4561	Rat long interspersed repetitive DNA	int
MMIRF12	353	Mouse interferon regulatory factor-2	opp CDS
MMMDM1A	1248 *	Mouse mdm-1 gene	
MMRPL3A	157	Mouse ribosomal protein 132	int near processed gene
MMTPMYOB	743	Mouse beta-tropomyosin	opp CDS
RNCYP45I	8878	Rat cytochrome P450IIE1;	
RNLCAG1	9687	Rat leukocyte common antigen;	opp CDS
RNLCAR	483	Rat leukocyte common antigen;	opp pot. glycosylation site
RNURIM	1292	Rat uricase	

Primates

GCGAL32	153	G.crassicaudatus short repeated DNA	int
HTLV1RES	589	Human HTLV-I related retroviral sequence;	int
HSARG1	162	Human arginase	
HSCALL01	5467	Human lymphoblastic leukemia antigen;	opp non-translated mR...
HSCN2	180	Human skin collagenase;	opp CDS
HSCN25	166	Human synovial collagenase	opp CDS
HSCOLLR	178	Human collagenase	
HSCYPJ	1613	Human cytochrome P-450j;	opp non-translated mRNA
HSENKPH2	140	Human enkephalin gene;	
HSFBRA	695	Human fibrinogen a-alpha-chain; HSFBRAA	opp CDS
HSFBRGG	8575	Human fibrinogen gamma chain; HSFBRGAB	
HSFIB1	1894 *	Human fibronectin;	int, alternative splicing!
HSGASTA	4231	Human gastrin gene;	int (near Alu)
HSGCRBR	2823	Human beta-glucocorticoid receptor;	
HSHLASBA	13320	Human HLA-SB(DP) alpha gene	int
HSHLASBA	12953	Human HLA-SB(DP) alpha gene	
HSMLCAB	900	Human alk. myosin light chain 1;	
HSMLC1F	900	Human alk. myosin light chain 1	
HSMLCAC	773	Human alkali myosin light chain 3;	opp non-coding mRNA
HSNMYC	6702	Human n-myc gene;	opp CDS
HSNMYC01	4949	Human n-myc gene;	opp intron
HSNMYC3	50	Human n-myc gene;	
HSOTC	282	ornithine transcarbamylase;	opp CDS
HSUG4PA	497 *	U4 small nuclear RNA pseudogene	

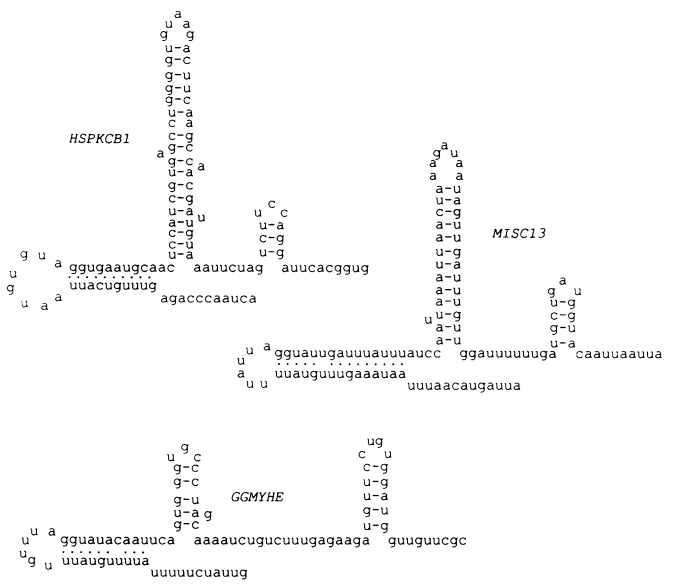

Figure 3. Three strong candidates for newly identified *trans*-splicing sites. The secondary structures of these predicted sites are very similar the known sites shown in Fig.1 HSPKCB1 is human protein kinase C, MISC13 is yeast mitochondrial, and GGMYHE is chicken embryonic myosin heavy chain. None of 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 rRNA, and *L. taraentolae* mitochondrial 12S 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.

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 mycoplasmas but this may well be due to the small amount of mycoplasma DNA in the database. Despite the availability of 38×10^6 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, HSPKCB1A, and HSTCVA5 contain the catalytic double guanosine [20] exactly at an exon boundary. The oxi3 locus in yeast (MISC13) has a well formed predicted *trans*-splicing site in the intron aI4. The intron is already known to be important for splicing [27] and the following self-splicing group II intron aI5g 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 (diverged *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 been detected. It is also possible that some putative sites are in fact pseudogenetic in nature. However, enough new

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 wide spread then 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., Wakasuigi, 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., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H. and Sugiura, 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 Suguira, 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., Shimada, H., Whittier, R., Ishibashi, 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 Sugiura, M. (1989) Mol. Gen. 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. Senpathy, 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. Tramontano, A., Scarlato, V., Barni, N., Cipollaro, M., Franze, A., Macchiato, M.F. and Cascino, A. (1984) Nucl. Acids Res. **12,** 5049–5059.
23. Kohchi, T.,Ogura, 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., Rindisbacher, 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 Slonoimski, P.P. (1982) Nature **298,** 628–632.
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.