Hypothesis

Selection of the 3′-splice site in group I introns

John M. Burke

Department of Microbiology, University of Vermont, Burlington, VT 05405, USA

Received 21 April 1989

A model for selection of 3′-splice sites in splicing of RNA precursors containing group I introns is presented. The key feature of this model is a newly identified tertiary interaction between the catalytic core of the intron and the 3′-splice site. This tertiary pairing would bring the 3′-splice site into the core of the intron, which is known to contain RNA sequences and structures essential for catalyzing the splicing reactions. The proposed tertiary interaction can coexist with P10, a pairing between 3′-exon sequences and the 'internal guide sequence' near the 5′-end of the intron. The model predicts that three RNA-RNA interactions are important in selection of 3′-splice sites: (i) binding of intron sequences with the core; (ii) pairing of exon sequences with the internal guide sequence; and (iii) binding of the terminal guanosine to an unknown site within the core.

RNA splicing; Intron, group I; Splice site; Ribozyme; RNA structure

In 1982, Davies and co-workers [1] used the sequences of nine group I introns to propose that both the 5′- and 3′-splice sites were selected by virtue of intramolecular base pairing with an 'internal guide sequence' (IGS) located near the 5′-end of the intron (fig.1). Pairing of the 5′-splice site with the IGS is termed P1, while pairing of exon sequences adjacent to the 3′-splice site with the IGS is termed P10 [2]. Since that time, sequences of many more group I introns have been reported from phylogenetically diverse genomes, including nuclear, mitochondrial (mt), chloroplast and bacteriophage examples. Comparative analysis shows that although there is little conservation of sequence, the sequences vary in such a way that P1 and P10 base pairing is nearly always maintained [3,4]. Such covariation is taken as strong evidence that base pairing exists and is functionally important in vivo [5].

Correspondence address: J.M. Burke, Department of Microbiology, University of Vermont, Burlington, VT 05405, USA

Analysis of mutants and second-site revertants has proven that P1 is necessary for splicing of the *Saccharomyces cerevisiae* mt COB4 intron in vivo [6,7], and for self-splicing of the *Tetrahymena thermophila* nuclear LSU rRNA intron in vitro [8,9]. No comparable evidence exists to support P10. Deletion of P10[5′] (the portion of the IGS proposed to pair with the 3′ exon) did not prevent proper utilization of 3′-splice sites, demonstrating that P10 is neither necessary nor sufficient for correct 3′-splice site utilization in self-splicing of *Tetrahymena* LSU [10]. Individual and compensatory two-base mutations designed to disrupt and restore base pairing of P10 in the same intron have very weak phenotypes [11]. On the basis of kinetic and mutational studies [12,13], Cech and co-workers have proposed that intron sequences adjacent to the 3′-terminal G of *Tetrahymena* LSU are recognized in splicing and intron cyclization, but the location of the sequences with which they would interact was not known.

While it is the only universally conserved base at the 3′-splice site, the 3′-terminal G cannot itself be

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50 © 1989 Federation of European Biochemical Societies

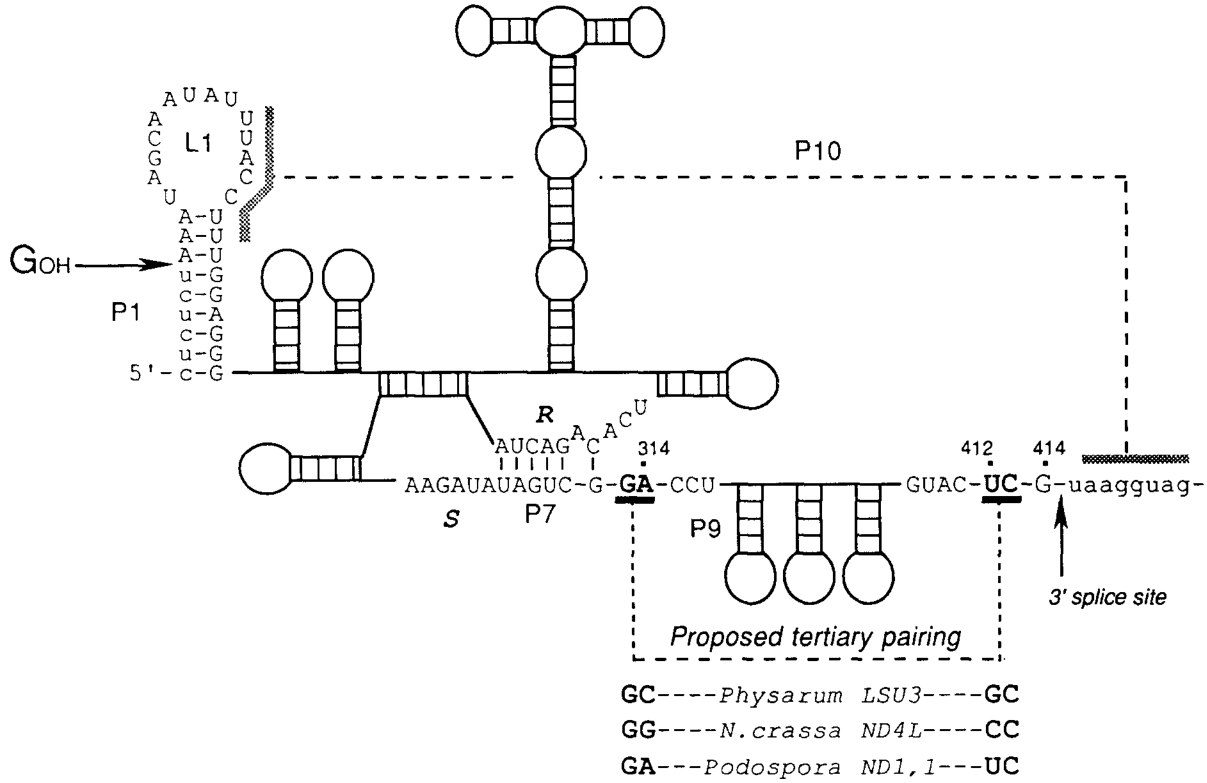


Fig.1. Intron structural model for *Tetrahymena* LSU. Proposed interactions involved in selection of the 3'-splice site (dotted lines) are superimposed on the secondary structure model [2]. Compensatory sequence variations at the site of the proposed tertiary pairing are shown for a subgroup of mt introns closely related to *Tetrahymena* LSU. R, S, P1, L1, P7, P9 and P10 are conserved sequences and structures characteristic of group I introns [2]. Uppercase, intron sequences. Lowercase, exon sequences.

sufficient to specify the 3'-splice site. Other Gs that are not used as splice sites are frequently found nearby. Protein factors may play a role in vivo, but 3'-splice site selection in self-splicing introns must be accomplished by structural features of the precursor RNA itself.

The *Physarum polycephalum* LSU3 intron is the closest known relative of *Tetrahymena* LSU. Except for a 0.5 kb insertion in *Physarum* LSU3, the sequences are 71% identical. More than half of the sequence variations are due to compensatory substitutions in established internal secondary structure elements. The remaining sites were examined for covariation consistent with base pairing, and this was observed between sites corresponding to positions 314 and 412 of *Tetrahymena* LSU (fig.1). Extension of this analysis to two closely related mt introns [14,15] revealed a further compensatory change (fig.1). These compensatory substitutions provide evidence for pairing between positions 314 and 412 in these four introns. Adjacent bases (G313 and C413) are also complementary, although invariant in this intron subgroup. Thus, two base pairs might form between the bases immediately downstream of the conserved sequence element S (313–314) and those preceding the terminal G of the intron (412–413).

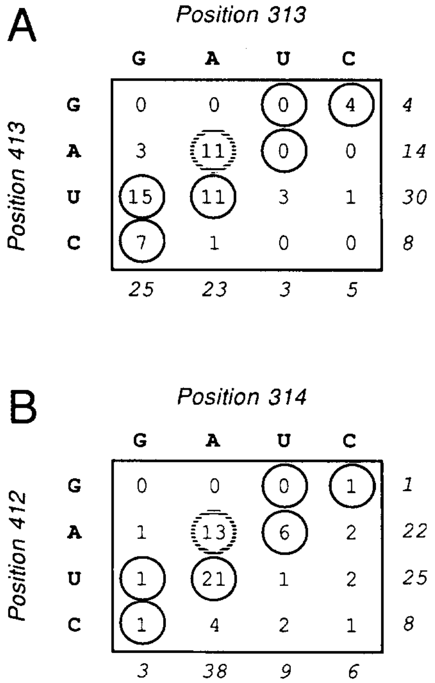
Sequences of these sites in all available group I introns were surveyed (figs 2 and 3). Positions 313 and 314 are strongly biased towards purines. Position 313 is a purine in 86% of the introns, while 68% of the bases at position 314 are A. Although there is no sequence conservation at positions 412 and 413, there exists a strong bias against G [4]. Strikingly, in each of the four cases where G is present at 413, a C is found as its putative pairing partner at 313 (fig.2A). Similarly, in the single case where G is present at 412, a C is found at 314 (fig.2B).

At positions 313–413, a Watson-Crick (G·C or A·U) or a G·U pair can form in 66% of the introns examined (fig.2A). The most common pair at this site is G·U, followed by A·U and A·A, G·C and C·G. A313·A414 pairs are much more common than any other nonstandard pairing at this site. Symmetrical A·A pairs involving N6 and N7 of both adenines are known [16] and have been proposed to occur at another site (A270·A304) in many group I introns [17]. Together, Watson-Crick, G·U and A·A pairs make up 86% of the 313-413 pairs in the introns surveyed.

At positions 314–412, a Watson-Crick or G·U pair can form in 54% of the introns examined (fig.2B). A·U pairs predominate (21 introns). A·A pairs are again by far the most frequent nonstandard pairing (13 introns). Watson-Crick, G·U and A·A pairs make up 77% of the 314–412 pairs.

When all four sites are considered together (fig.3), 21 introns (38%) can form two Watson-Crick or G·U base-pairs between positions 313–413 and 314–412. Of the remaining 35 introns, 25 can form one base-pair and 11 of these have A·A pairs at the other site. Among the ten introns that cannot form a Watson-Crick or G·U pair at either site, five can form two A·A pairs and three can form one A·A pair. Only two introns, *S. pombe* mt OX3 and *A. nidulans* mt OX1, cannot form Watson-Crick, G·U or A·A pairs at either site.

The proposed tertiary pairing is supported by results of mutations at position 413 of *Tetrahymena* LSU [13]. The mutation C413A results in a 40% reduction in self-splicing activity, but does not affect 3′-splice site selection. In the proposed tertiary pairing, G313-C413 is replaced by G313-A413 (fig.4), a change that is expected to destabilize but not eliminate the tertiary pairing, since G-A pairs at these positions are found in three other introns (fig.2A). However, the mutation C413G results in a switch in 3′-splice site selection, causing exon ligation to follow G413 instead of G414. As indicated above, G is observed to precede the 3′-terminal G only when its pairing partner in the core is C. The model suggests that in the C413G mutant an alternative pairing forms, in which G313 and A314 pair with C411 and U412 (fig.4). This altered alignment results in exon ligation following G413 rather than G414. A prediction of the model is that the mutation G313C would restore exon ligation to the phosphodiester bond following G414 in the C413G mutant.



A

Position 313

Position 413	G	A	U	C	
G	0	0	0	4	4
A	3	11	0	0	14
U	15	11	3	1	30
C	7	1	0	0	8
	25	23	3	5	

B

Position 314

Position 412	G	A	U	C	
G	0	0	0	1	1
A	1	13	6	2	22
U	1	21	1	2	25
C	1	4	2	1	8
	3	38	9	6	

Fig.2. Covariation of sequence at sites of proposed tertiary pairing in 56 group I introns. (A) Sequences at position corresponding to G313·C413 of *Tetrahymena* LSU. (B) Sequences at position corresponding to A314·U412 of *Tetrahymena* LSU. Solid circles, A·U, G·C and G·U pairs. Dotted circles, A·A pairs.

The proposed tertiary pairing between the core (positions 313 and 314) and intron sequences at the 3′-splice site (positions 412 and 413) may coexist with P10, the 3′-splice site–IGS pairing (figs 1 and 5). While pairing between the 5′-splice site and the IGS involves sequences which span the splice site, pairing between the 3′-splice site and the IGS is proposed to involve only the exon sequences immediately adjacent to the splice site [3]. Thus, intron sequences at the 3′-splice site are available to participate in other interactions, such as the proposed tertiary pairing.

Kinetic evidence shows that the terminal G of *Tetrahymena* LSU is specifically recognized and bound by an unknown site within the IVS [12,18]. Mutations at this site (G414U, G414C) greatly

Two base-pairs

An	LSU	**GA**----**UU**G/
Cm	LSU1	**GA**----**UU**G/
Cm	LSU2	**GA**----**UU**G/
Cm	LSU5	**AA**----**UU**G/
Cm	SSU	**GA**----**UC**G/
Cr	LSU	**AU**----**AU**G/
Kt	LSU	**GA**----**UU**G/
Mp	tRNAL	**GU**----**AU**G/
Nc	COB2	**AA**----**UU**G/
Nc	LSU	**GA**----**UU**G/
Nc	ND1	**AA**----**UU**G/
Nc	ND4L	**GG**----**CC**G/
Nc	ND5,2	**GA**----**UU**G/
Pa	ND1,1	**GA**----**UC**G/
Pa	OX7	**AU**----**AU**G/
Pp	LSU3	**GC**----**GC**G/
Sc	COB5	**GA**----**UU**G/
Sc	LSU	**GA**----**UU**G/
Sp	OX2a	**AA**----**UU**G/
T4	sunY	**GA**----**UU**G/
Tt	LSU	**GA**----**UC**G/

One base pair

An	COB	**G**A----A**U**G/
Nc	ND3	**G**A----A**C**G/
Nc	OX2	**A**A----A**U**G/
Sc	OX3	**A**A----A**U**G/
Sc	OX5b	**A**A----A**U**G/
Zm	tRNAL	**C**A----A**G**G/
An	OX2	A**A**----**U**AG/
Pa	ND1,2	A**U**----**A**AG/
Pa	OX9	A**U**----**A**AG/
Sc	COB4	A**U**----**A**AG/
SPO1	g31	A**G**----**U**AG/
Ce	LSU5	**G**C----U**U**G/
Cm	LSU3	**G**A----C**U**G/
Cm	LSU4	**G**A----C**U**G/
Nc	ATP6,2	**A**C----C**U**G/
Pa	OX8	**A**C----U**U**G/
Pp	LSU1	**C**U----C**G**G/
Pp	LSU2	**C**U----C**G**G/
T4	nrdB	**G**G----A**C**G/
T4	td	**G**C----A**U**G/
Vf	tRNAL	**C**A----C**G**G/
An	OX3	U**A**----**U**UG/
Nc	OX3	C**A**----**U**UG/
Nc	OX4	U**A**----**U**UG/
Sp	OX2	U**A**----**U**UG/

No base pairs

Cm	psbA1	AA----AAG/
Kf	ATP9	AA----AAG/
Nc	ND5,1	AA----AAG/
Sc	OX4	AA----AAG/
Sc	OX5a	AA----AAG/
Cm	psbA2	AC----AAG/
Nc	COB1	AA----ACG/
Sc	COB3	GA----AAG/
Sp	OX3	GA----CAG/
An	OX1	GU----UAG/

Fig.3. Sequences at positions of proposed tertiary pairing in 56 group I introns. Sequences shown are at positions corresponding to (from left to right) G313, A314, U412, C413 and G414 of *Tetrahymena* LSU. Slashes mark 3′-splice sites. Boldface letters mark A·U, G·C and G·U pairs. Underlining, A·A pairs. Standard abbreviations for group I introns [4] are used. To save space, names of organisms, introns, and references to sequences are not included here; all can be found in the review by Cech [4].

reduce the rate of exon ligation [13]. Strikingly, exon ligation proceeds at the normal site in these mutants cryptic 3′-splice sites are not activated. This indicates that while a G at the 3′-end of the

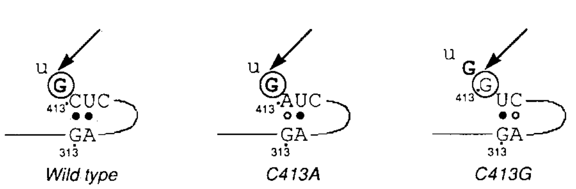


Fig.4. Effect of mutations at C413 [13] on 3′-splice site selection in *Tetrahymena* LSU. Intron sequences are indicated by uppercase letters; the lowercase U is the first base of 3′-exon. Boldface, G414. Circle indicates binding site for base at 3′-end of the intron. Arrow indicates 3′-splice site. For further discussion, see text.

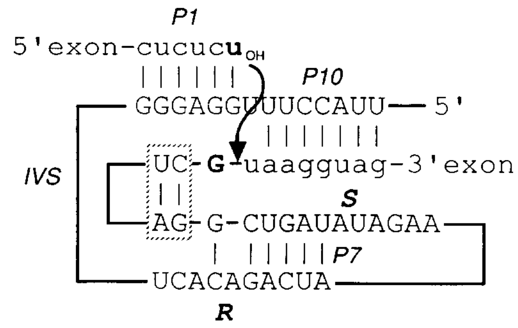


Fig.5. Model for 3′-splice site structure during exon ligation. The sequences and structures shown are those proposed for *Tetrahymena* LSU, but would be similar for most group I introns. Uppercase, intron sequences. Lowercase, exon sequences. Proposed tertiary pairing is indicated by box.

intron is important for efficient exon ligation, it is not an essential determinant of the 3′-splice site in self-splicing of *Tetrahymena* LSU.

In summary, I have presented a model in which three distinct RNA-RNA interactions participate in selection and utilization of 3′-splice sites in splicing of group I introns (fig.5). First, intron sequences immediately adjacent to the 3′-terminal G may form a tertiary pairing with complementary sequences immediately downstream of core sequence element S. Second, the 3′-terminal G is recognized by a site whose location is unknown, but is likely to be within the highly conserved sequences of the intron core [4]. Third, exon sequences immediately adjacent to the 3′-terminal G may pair with an internal guide sequence near the 5′-end of the intron. Since these three interactions involve adjacent bases spanning the 3′-splice site, the three binding sites must be in close physical proximity. This places significant constraints on possible three-dimensional models of group I intron structure. For example, the binding site for the terminal G might lie within the highly conserved bases at the 5′-end of element R (fig.5).

In addition to splicing, the proposed tertiary pairing may be important for a number of other reactions involving the terminal G of *Tetrahymena* LSU, including intron cyclization, circle reopening with CpU, *trans*-splicing and hydrolysis reactions. The model can be tested by mutational and modification-protection experiments, which are currently underway.

Acknowledgements: I wish to thank R. Collins, V. Vogt, C. Lemieux and D. Shub for communicating unpublished sequence data. This work was supported by NIH grant GM36981.

REFERENCES

[1] Davies, R.W., Waring, R.B., Ray, J.A., Brown, T.A. and Scazzocchio, C. (1984) Nature 300, 719–724.

[2] Burke, J.M., Belfort, M., Cech, T.R., Davies, R.W., Schweyen, R.J., Shub, D.A., Szostak, J.W. and Tabak, H.F. (1987) Nucleic Acids Res. 15, 7217–7221.

[3] Waring, R.B., Davies, R.W., Brown, T.A. and Scazzocchio, C. (1984) Gene 28, 277–291.

[4] Cech, T.R. (1988) Gene 73, 259–271.

[5] Noller, H.F. (1984) Annu. Rev. Biochem. 53, 119–162.

[6] De La Salle, H., Jacq, C. and Slonimski, P.P. (1982) Cell 28, 721–732.

[7] Perea, J. and Jacq, C. (1985) EMBO J. 4, 3281–3288.

[8] Waring, R.B., Towner, P., Minter, S.J. and Davies, R.W. (1986) Nature 321, 133–139.

[9] Been, M.D. and Cech, T.R. (1986) Cell 47, 207–216.

[10] Been, M.D. and Cech, T.R. (1985) Nucleic Acids Res. 13, 8389–8408.

[11] Davies, R.W., Waring, R.B. and Towner, P. (1987) Cold Spring Harbor Symp. Quant. Biol. 52, 165–172.

[12] Tanner, N.K. and Cech, T.R. (1987) Biochemistry 26, 3330–3340.

[13] Price, J.V. and Cech, T.R. (1988) Genes Dev. 2, 1439–1447.

[14] Michel, F. and Cummings, D.J. (1985) Curr. Genet. 10, 69–79.

[15] Nelson, M.A. and Macino, G. (1987) Mol. Gen. Genet. 206, 318–325.

[16] Saenger, W. (1983) Principles of Nucleic Acid Structure, Springer-Verlag, New York.

[17] Kim, S.-H. and Cech, T.R. (1987) Proc. Natl. Acad. Sci. USA 84, 8788–8792.

[18] Kay, P.S., Menzel, P. and Inoue, T. (1988) EMBO J. 7, 3531–3537.