Evidence of natural selection to maintain a functional domain outside of the 'core' in a large subclass of Group I introns

Richard A.Collins

Department of Botany, University of Toronto, Toronto, Ontario, Canada

Received October 22, 1987; Revised and Accepted February 15, 1988

ABSTRACT

Comparison of three closely-related, homologous Group I introns reveals conservation of RNA secondary structure and some primary sequence outside of the characteristic Group I core structure. Further examination of forty Group I introns showed that all can be placed into one of two categories based on the length of the "loop L5" region (subtended by the base-paired sequences P and Q): short (21 to 38 bases) or long (59 to 295 bases). Despite the large variation in size and sequence, all nineteen of the long L5 introns share a common structure whose features include an adenine-rich bulge at a fixed distance from the P-Q pairing. This bulge is flanked by base-paired regions of $\geq$ 6 base pairs on the core-proximal side and $\geq$ 3 base pairs on the distal side. In the core-proximal helix there are a large number and high proportion of deviations from the consensus sequence that maintain base-pairing. These naturally-occurring compensatory base substitutions provide compelling phylogenetic support for the existence of this pairing and indicate that the conserved structure has a function in vivo.

INTRODUCTION

The majority of introns in mitochondrial genes and nuclear rRNA genes are classified as Group I introns based on the presence and location of somewhat conserved sequences that allow the formation of a common "core" secondary structure (1-5). Primary sequence varies among introns, even in the base-paired sequences involved in formation of the core structure. Naturally-occurring deviations from the consensus in one sequence of a base-paired stem are almost always accompanied by the compensating base substitution in the other sequence such that the secondary structure is maintained. Perhaps the most interesting consequence of these intron structures is that they confer upon at least some introns the ability to self-splice in vitro in the absence of proteins (reviewed by Cech (6)).

The sizes of naturally-occurring Group I introns range from 258 bases to over 2 kilobases (4,7,8). While the larger sizes often reflect the presence of an intron-encoded open reading frame, many introns also contain non-coding sequences outside of the core. The importance of these "non-core" sequences is not known. Such non-core sequences have received very little attention, since it is often suggested that all Group I introns splice via the same basic mechanism, involving only sequences and secondary structures contained in the core. However, it has been noted that some of the sequences outside of the core in a few mitochondrial introns show some similarity to the corresponding regions of the intron of the *Tetrahymena* rRNA gene (2,8,9). These similarities have been taken to support the view that Group I introns are descended from a common ancestor (10,11).

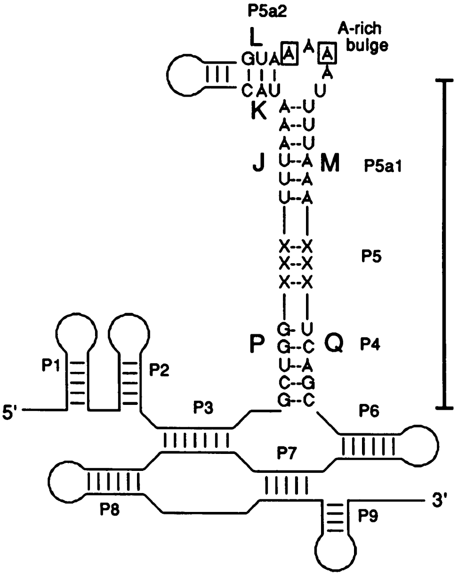
To search for functionally important regions outside of the core of Group I introns, and to describe the molecular events involved in evolution of Group I introns, we have been analyzing closely-related homologous introns. These comparisons have revealed that certain sequences outside of the core structure appear to be subject to natural selection that has maintained RNA secondary structure and in some cases primary sequence, implying a function for these sequences *in vivo*.

MATERIALS AND METHODS

The secondary structures presented here were constructed by examining primary sequence, potential base pairing, and the phylogenetic occurrence of compensating base substitutions that would maintain secondary structure in Group I introns. Similar approaches have previously been used to predict Group I and II intron core structures (1-4) and secondary structures in ribosomal RNAs (12). Computer programs and an IBM PC/AT were used to assist in identification of complementary sequences (International Biotechnologies Inc.; see (13)) and potential secondary structures (PCFOLD version 3.0, M. Zuker, National Research Council of Canada, Ottawa, Ontario; see (17)).

RESULTS and DISCUSSION

Aspergillus nidulans and *Schizosaccharomyces pombe* contain similar Group I introns at precisely the same location in their mitochondrial *COI* gene (14,15). In some natural isolates of *Neurospora crassa* a similar intron is also present at this position (Collins, unpublished data). Comparison of these three homologous intron sequences reveals several insertions/deletions. Four of the additional sequences in *S. pombe* that are not present in *Neurospora* or *Aspergillus* occur as two complementary pairs in the variable length "loop" subtended by the base paired regions P4 (P-Q) and P5 (Figure 2A, B, I; see Figure 1 for a structural diagram of a Group I intron). This region is designated loop L5 in the recently-proposed Group I intron nomenclature



<u>Figure 1</u>. Potential secondary structure of a Group I intron. Base-paired regions P1 to P9 and relevant sequences involved in pairing, P, Q, R, S, J, K, L, M, are indicated. The consensus sequence and possible secondary structure of introns with a long loop L5 region (subtended by the helices P4 and P5) is emphasized. The drawing is fashioned after the recently-proposed structural conventions for Group I introns (5). In most introns with a short L5 region, the P5 pairing closes only a small loop (see Figure 1A in ref. 5). In accordance with the proposed nomenclature, the additional helix subtended by the P5 pairing in all of the introns with a long L5 region is designated P5a. The portions of helix P5a on the core-proximal and -distal sides of the adenine-rich bulge are designated P5a1 and P5a2, respectively, to facilitate discussion; these portions are formed by base pairing of sequence J with M, and K with L, respectively (see text). The bar indicates the constant distance between the P-Q pairing and the conserved, unpaired adenines (boxed) in the bulge. The non-conserved sequences that comprise the P5 pairing (1,2) are indicated as X. See Figures 2 and 3 for sequences of individual introns.

conventions (5); it has previously been called loop L3 (1) or the C-helix (3). The occurrence of complementary pairs of insertions suggests evolutionary conservation of a secondary structure. This observation, coupled with previously noted similarities among parts of the L5 regions of the *Tetrahymena* rRNA intron (16) and some mt introns in the *Neurospora* *URF5* (9), *Podospora* *URF1* (8) and *Saccharomyces* *OXI3* and *COB* (2) genes, prompted us to analyze this region in other Group I introns.

A Model for the Structure of the L5 Region of Group I Introns.

The 40 Group I introns examined could be placed into one of two clearly distinguishable categories based on the distance between the conserved sequences P and Q that form the P4 helix. In 21 of the introns this region is short (21 to 38 nt) and can form a single helix (P5) of variable length. P5 often includes some unpaired bases and encloses a variable size loop (L5) at the end. In the other 19 introns P and Q are separated by a much greater distance (59 to 295 nt).

A detailed examination (Figures 2 and 3) reveals that all 19 of the introns with the long L5 region share a common structure (Figure 1) which has the following features:

1) five unpaired bases forming an "adenine-rich" bulge at a fairly constant distance from the P-Q pairing;
2) bases 1 and 3 of the bulge are invariably adenine;
3) the bulge is flanked by helices of $\geq$ 6 bp (P5a1) on the core-proximal side and $\geq$ 3 bp (P5a2) on the core-distal side;
4) additional sequences in individual introns form bulges or stem-loops extending from one or more of a limited number of positions in the conserved structure. These peripheral structures have little, if any, effect on the distance between the A-rich bulge and the P-Q pairing.

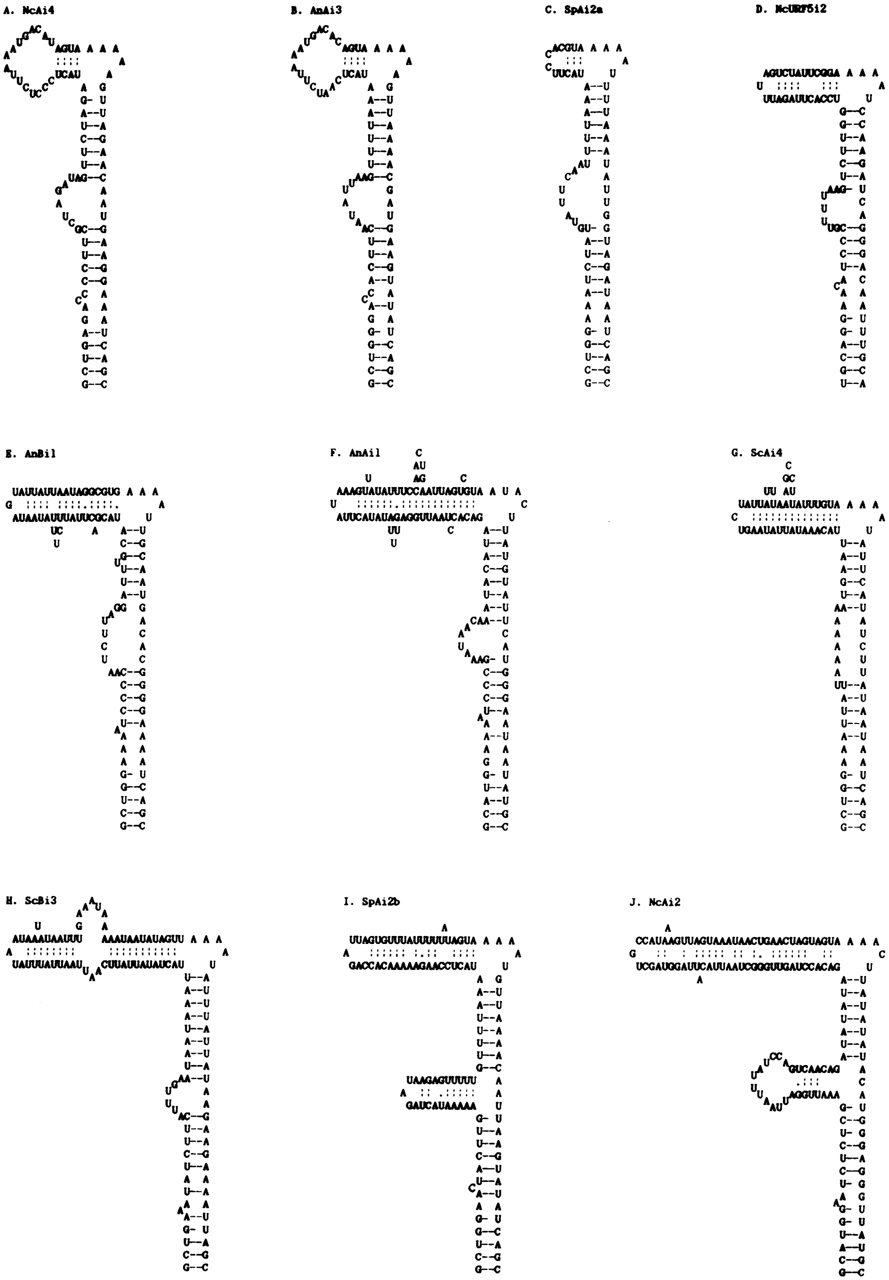
Sequences J and M, comprising the P5a1 pairing, show many deviations from the consensus (Figure 3). Of 228 nucleotides (19 introns, 2 strands per helix, 6 nucleotides per strand), only 131 match the consensus. Ninety-three of the 97 deviations are substitutions that maintain base pairing. This large number and high proportion of deviations that maintain base pairing provide compelling phylogenetic support for the existence of the P5a1 pairing. Of the four deviations that disrupt pairing, three likely reflect a single ancestral substitution, since they are found at the same position in homologous introns located at the same position (base 1 in the M sequence) in the _coI_ genes of _Neurospora_, _Aspergillus_, and _Schizosaccharomyces_ (Figure 3 A, B, I). In these three introns one additional base pair (G-C) is present at the core-proximal side of P5a1, restoring the number of base pairs to six (Figure 2 A-C). The other disruption occurs in the _Physarum_ nuclear rRNA intron, where an A-U is replaced by an A-A. The weakening of this helix by the unpaired adenines is likely counteracted by the surrounding paired bases, five of which are G-C pairs rather than the consensus A-U pairs (Figure 2 S).

In contrast to the large number of substutitions in the sequences in the P5a1 helix, the sequences K and L which pair to form P5a2 are well conserved: of 114 nucleotides, 103 match the consensus (Figure 3). The small number of compensating substitutions makes it more difficult to evaluate the importance of base pairing in this region. Indeed, 6 of the 11 substitutions disrupt one of these three base pairs, although substantial pairing of distal sequences may offset this (Figure 2). Nonetheless, there seems to be some selection for primary sequence in the nucleotides that form the P5a2 pairing.

The diagrams in Figure 1 and 2 are drawn to emphasize the features of the secondary structure model presented here. Neither the three dimensional structure of this region nor of the intron core has yet been established for any intron. Deciphering the in vivo structure(s) of these introns could be very complicated, especially for introns in which trans-acting proteins play an important role in determining RNA structure. The structures shown in Figure 2 as well as those usually drawn for the core of all Group I introns (1-5) probably do not represent the most stable structure of the free intron RNA, but possibly a transition-state or one of several alternative conformations assumed by the intron at a particular stage during splicing. Indeed, using a computer program (PCFOLD version 3.0; see (17)) to calculate lowest free energy structures of the L5 regions, the structures shown in Figure 2 were obtained for only about 13 of the 19 introns (not shown). Thus, comparative sequence analysis and mutation/suppression experiments may be more informative approaches to the initial identification of additional functionally-important regions within Group I intron RNAs.

Another sub-group of Group I introns has been recognized previously. These Group IA introns (2) contain additional sequences between P7 and P3. The presence of these additional sequences does not correlate perfectly with a short vs. long L5 region: although most of the Group IA introns have a short L5, at least one, ScBi3 (Figure 2H), has a long L5.

Thusfar, no common functional property of all the long L5 introns has been recognized. The only completely conserved feature of the structural model presented here is the adenine-rich bulge flanked by base-paired regions, at least one of which is not well conserved in primary sequence. This may mean that the purpose of the paired bases is simply to provide a bulged adenine(s) at a particular position relative to some other structural feature in the intron. If the unpaired adenine(s) are the important feature of the Group I intron L5 structure, some of the 'short L5' introns may also fit the model. Many, although not all, of the introns with a short L5 have one or more unpaired adenines in the loop at the end of the P5 pairing (4; Collins, unpublished). In most cases, because of the absence of the P5a1 pairing, these adenines would probably be closer to the P-Q pairing than those in the long L5 introns, but may serve the same function.


A. NcAi4
B. AnAi3
C. SpAi2a
D. NcURF6i2
E. AnBi1
F. AnAi1
G. ScAi4
H. ScBi3
I. SpAi2b
J. NcAi2

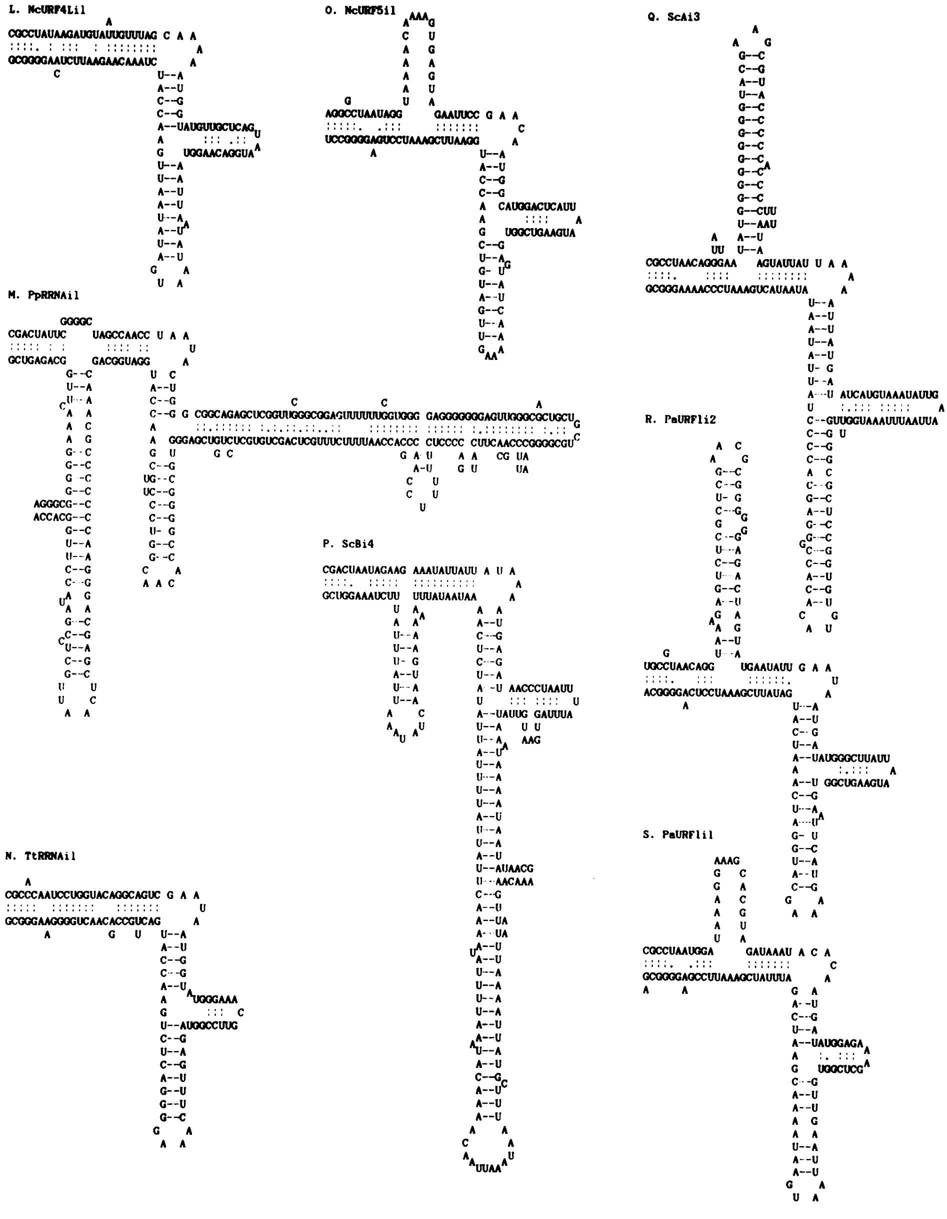


Figure 2. Secondary structures of all Group I introns with long L5 regions. Only the portion of each intron corresponding to the consensus structure emphasized in Figure 1 (beginning at the paired regions of sequences P and Q) is shown. Orientation is the same as in Figure 1. See Figure 3 for abbreviations used and references.

		5' J	K		L	A-RICH BULGE	M 3'	REFERENCES
A.	NcAi4	UCUAGA	UAC	18 nucleotides	GUA	AAAAA	GUUAGA	Collins, unpublished
B.	AnAi3	UUUAAA	UAC	18	GUA	AAAAA	GUUAAA	14
C.	SpAi2a	UUUAAA	UAC	6	GUA	AAAAU	UUUAAA	7
D.	NcURF5i2	UCUUGG	UCC	19	GGA	AAAAU	CCAAGA	9
E.	AnBi1	U AUUGCA	UAC	33	GUG	AAAAU	UGCAAU	38
F.	AnAi1	UACAUA	GAC	52	GUA	AUACU	UAUGUA	14
G.	ScAi4	AUGAAU	UAC	34	GUA	AAAAU	AUUCAU	40
H.	ScBi3	AUAAAU	UAC	54	GUU	AAAAU	AUUUAU	42
I.	SpAi2b	UUUAAA	UAC	36	GUA	AAAAU	GUUAAA	15
J.	NcAi2	UAUAUA	GAC	60	GUA	AAACU	UAUAUA	Collins, unpublished
K.	NcURF4Li1	CAAAUC	UAC	51	GUA	AAAAC	GAUUUG	9
L.	TtRRNAi1	U CGUCAG	UAC	43	GUA	AUAAG	CUGACG	36
M.	NcURF5i1	UUAAGG	UAC	50	GUA	ACAAG	CCUUAA	9
N.	PaURF1i1	UAUUUA	GAC	42	GUA	ACACA	UAAAUA	8
O.	ScBi4	UAAUAA	AAC	126	GUA	AAAUA	UUAUUA	41
P.	NcATP6i2	UUUAGG	UAC	9	GUA	AAAAU	CUUAAA	39
Q.	PaURF1i2	UUAUAG	UAC	50	GUA	AUAAG	UUAUAA	8
R.	ScAi3	AUAAUA	UAA	79	UUA	AAAAU	UAUUAU	40
S.	PpRRNAi1	GGUAGG	UAC	175	GUC	AUAAU	CCAACC	37

	J						K			L			A-RICH BULGE					M					
U	12	10	11	4	5	2	15	0	0	1	18	1	0	4	0	1	11	8	10	12	6	5	4
C	2	2	1	1	1	1	0	1	18	0	0	1	0	2	0	3	1	5	3	1	1	2	1
A	4	5	6	13	8	10	1	18	1	0	0	16	19	13	19	15	4	2	5	5	11	10	13
G	1	2	1	1	5	6	3	0	0	18	1	1	0	0	0	0	3	4	1	1	1	2	2
CONSENSUS:	U	U	U	A	A	A	U	A	C	G	U	A	A	A	A	A	U	U	U	U	A	A	A
Deviations from the consensus that:																							
1) maintain pairing	7	9	8	6	11	9	0	1	1	1	1	1						8	9	6	8	9	7
2) disrupt pairing	0	0	0	0	0	0	4	0	0	0	0	2						3	0	1	0	0	0

Figure 3. Summary of the sequences that form the consensus secondary structure in the long L5 region of Group I introns. Introns are designated in a manner similar to that used by Waring and Davies (4). The first two letters represent the genus and species (Tt: Tetrahymena thermophila; Pp: Physarum polycephalum; Nc: Neurospora crassa; Pa: Podospora anserina: An: Aspergillus nidulans; Sc: Saccharomyces cerevisiae; Sp: Schizosaccharomyces pombe). These are followed by the gene designation (A: subunit I of cytochrome c oxidase; B: apocytochrome b; RRNA: nuclear large rRNA; URFn: unassigned reading frame usually corresponding to the human URF (ND) gene of the same number, n; ATP6: subunit 6 of the F_1/F_0 ATPase). i1 through i4 refer to the intron number within a gene, starting at the 5' end. Bulged bases in the J sequence of TtRRNAi1 and AnBi1 are shown above the line. SpAi3 as drawn by Trinkl and Wolf (7) would also appear to have a long L5 region. However, this intron structure is unusual in having two stem/loops between sequences E and P. An alternative choice for E and P allows a more typical structure to be drawn and suggests that this intron

may have a short L5 region (Collins, data not shown). SpAi3 is therefore not included in the present analysis.

The number of occurrences of each base at each position in sequences J, K, L, bulge, and M are tabulated below the individual sequences. The consensus sequence consists of the sequence of the most-frequent base at each position. For the base-paired sequences, a table of the number of deviations from the consensus sequence that either maintain or disrupt pairing is presented at the bottom.

The bases in the adenine-rich bulge could be involved in interactions with other parts of the intron or with trans-acting factors. In other systems, such as prokaryotic ribosomes (reviewed in 18), mRNA leader sequences (19) and R17 bacteriophage RNA (20,21), bulged residues, particularly adenines, are important for specific protein-RNA interactions. The importance _in vivo_ of trans-acting proteins even in RNA-catalyzed reactions has already been demonstrated. For example, the _Neurospora_ apocytochrome b intron 1 self-splices _in vitro_ (22) but requires the nuclear _cyt18_ gene product for splicing _in vivo_ (23). Also, splicing of the intron in the _Neurospora_ mitochondrial large rRNA gene, which proceeds by the guanosine-initiated transesterification mechanism characteristic of self-splicing introns (24,6) requires the _cyt18_ gene product _in vivo_ (25) and does not occur _in vitro_ in the absence of added mitochondrial protein(s) (26). The protein(s) appear to be required for correct folding of the RNA (27). In yeast, nuclear (28,29) and mitochondrial (30-33) genes encode trans-acting factors required for splicing of individual mitochondrial Group I introns. The site(s) of interaction of trans-acting factors with the Group I introns is not yet known. The long L5 region of the _Neurospora_ introns is not likely to be an essential part of the binding site of the _cyt18_ gene product: several Group I introns with a short L5 region are nonetheless dependent on the _cyt18_ gene product for splicing _in vivo_ (23). However, the variations in L5 structure in different introns could provide intron-specific binding sites for maturases or other factors that are required for splicing of individual introns.

Conservation of the L5 structure in such a wide range of organisms indicates that the structure has a function _in vivo_. This structure might be involved in splicing or possibly in some other intron-mediated reaction. It will be possible to test by site-directed mutagenesis whether any parts of the model presented here are relevant to RNA-catalyzed splicing _in vitro_. It can be concluded _a priori_ that a long L5 region is not one of the minimal structural requirements for RNA-catalyzed splicing, since at least one self-splicing intron, intron 1 in the _Neurospora_ apocytochrome b gene (22) has a short L5 region (34,35). It is possible that other sequences or structures substitute for the long L5 region in self-splicing introns that have a short L5 region. Alternatively, the L5 region might not be essential for splicing *per se*, but may influence the choice of splice sites or the rate of reaction, either by affecting RNA structure directly or by supplying a binding site for trans-acting factors.

ACKNOWLEDGEMENTS

I thank and Drs. Deborah Field and Shane Climie for comments on the manuscript. This work was supported by grants from the National Research Council of Canada and the University of Toronto.

REFERENCES

1. Davies,R.W., Waring,R.B., Ray,J.A., Brown,T.A. and Scazzocchio,C. (1982) Nature 300,719-724.
2. Michel,F., Jacquier,A. and Dujon,B. (1982) Biochimie 64,867-881.
3. Michel,F. and Dujon,B. (1983) EMBO J. 2,33-38.
4. Waring,R.B. and Davies,R.W. (1984) Gene 28,277-291.
5. 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) Nucl. Acids Res. 15,7217-7221.
6. Cech,T.R. (1986) Cell 44,207-210.
7. Trinkl,H. and Wolf,K. (1986) Gene 45,289-297.
8. Michel,F. and Cummings,D.J. (1985) Curr. Genet. 10,69-79.
9. Nelson,M.A. and Macino,G. (1987) Mol. Gen. Genet. 206,318-325.
10. Borst,P. and Grivell,L.A. (1981) Nature 289,439-440.
11. Hensgens,L.A.M., Bonen,L., de Haan,M., van der Horst,G. and Grivell,L.A. (1983) Cell 32,379-389.
12. Noller,H., Kop,J., Wheaton,V., Brosius,J., Gutell,R.R., Kopylov,A.M., Dohme,F., Herr,W., Stahl,D.A., Gupta,R. and Woese,C. (1981) Nucl. Acids Res. 9,6167-6189.
13. Pustell,J. and Kafatos,F.C. (1984) Nucl. Acids Res. 12,643-655.
14. Waring,R.B., Brown,T.A., Ray,J.A., Scazzocchio,C. and Davies,R.W. (1984) EMBO J. 3,2121-2128.
15. Lang,B.F. (1984) EMBO J. 3,2129-2136.
16. Cech,T.R., Tanner,N.K., Tinoco,I.Jr., Weir,B.R., Zuker,M. and Perlman, P.S. (1983) Proc. Natl. Acad. Sci. USA 80,3903-3907.
17. Zuker,M. and Stiegler,P. (1981) Nucl. Acids Res. 9,133-148.
18. Garrett,R.A., Vester,B., Leffers,H., Sorensen,P.M., Kjems,J., Olesen,C.A., Christiansen,J. and Douthwaite,S. (1984) In Clark, B.F.C. and Petersen,H.U. (eds.), Gene Expression, Alfred Benzon Symposium 19. Munksgaard, Copenhagen. pp.331-352.
19. Climie,S. and Friesen,J.D. (1987) J. Mol. Biol. 198,371-381.
20. Carey,J., Lowary,P.T. and Uhlenbeck,O.C. (1983) Biochemistry 22,4723-4730.
21. Romaniuk,P.J., Lowary,P., Wu,H-N., Stormo,G. and Uhlenbeck,O.C. (1987) Biochemistry 26,1563-1568.
22. Garriga,G. and Lambowitz,A.M. (1984) Cell 38,631-641.
23. Collins,R.A. and Lambowitz,A.M. (1985) J. Mol. Biol. 184,413-428.
24. Garriga,G. and Lambowitz,A.M. (1983) J. Biol. Chem. 258,14745-14748.
25. Mannella,C.A., Collins,R.A., Green,M.R. and Lambowitz,A.M. (1979) Proc. Natl. Acad. Sci. USA 76,2635-2639.
26. Garriga,G. and Lambowitz,A.M. (1986) Cell 46,669-680.
27. Wollenzien,P.L., Cantor,C.R., Grant,D.M. and Lambowitz,A.M. (1983) Cell 32,397-407.
28. McGraw,P. and Tzagoloff,A. (1983) J. Biol. Chem. 258,9459-9468.

29. Kreike,J., Schulze,M., Pillar,T., Korte,A. and Rodel,G. (1986) Curr. Genet. 11,185-191.
30. Lazowska,J., Jacq,C. and Slonimski,P.P. (1980) Cell 22,333-348.
31. Weiss-Brummer,B., Rodel,G., Schweyen,R.J. and Kaudewitz,F. (1982) Cell 29,527-536.
32. Anziano,P.Q., Hanson,D.J., Mahler,H.R. and Perlman,P.S. (1982) Cell 30,925-932.
33. De La Salle,H., Jacq,C. and Slonimski,P.P. (1982) Cell 28,721-732.
34. Helmer-Citterich,M., Morelli,G. and Macino,G. (1983) EMBO J. 2,1235-1242.
35. Burke,J.M., Breitenberger,C., Heckman,J.E., Dujon,B. and RajBhandary,U.L. (1984) J. Biol. Chem. 259,504-511.
36. Kan,M.C. and Gall,J.G. (1982) Nucl. Acids Res. 10,2809-2822.
37. Nomiyama,H., Sakaki,Y., and Takagi,Y. (1981) Proc. Natl. Acad. Sci. USA 78,1376-1380.
38. Waring,R.B., Davies,R.W., Scazzocchio,C. and Brown,T.A. (1982) Proc. Natl. Acad. Sci. USA 79,6332-6336.
39. Morelli,G. and Macino,G. (1984) J. Mol. Biol. 178,491-507.
40. Bonitz,S.G., Coruzzi,G., Thalenfeld,B.E., Tzagoloff,A. and Macino,G. (1980) J. Biol. Chem. 255,11927-11941.
41. Nobrega,F. and Tzagoloff,A. (1980) J. Biol. Chem. 255,9828-9837.
42. Holl,J., Schmidt,C. and Schweyen,R.J. (1985) In E. Quagliariello et al. (eds.), Achievements and Perspectives of Mitochondrial Research. Vol. II. Elsevier Science Publishers, Amsterdam. pp.227-237.