Proc. Natl. Acad. Sci. USA
Vol. 77, No. 4, pp. 1877–1879, April 1980
Biochemistry

A mechanism for RNA splicing

(small nuclear RNA/RNA splice sites/intervening sequences/RNA processing)

JOHN ROGERS AND RANDOLPH WALL

Molecular Biology Institute and Department of Microbiology and Immunology, UCLA School of Medicine, Los Angeles, California 90024

Communicated by P. D. Boyer, January 2, 1980

ABSTRACT **The most abundant of the stable small nuclear RNAs of eukaryotic cells, U-1 small nuclear RNA, is exactly complementary to the consensus sequences at RNA splice sites. We propose that this RNA is the recognition component of the nuclear RNA splicing enzyme and forms base pairs with both ends of an intron so as to align them for cutting and splicing.**

Intramolecular RNA splicing, first discovered for the adenovirus-specific RNA (1–3), occurs in the expression of numerous eukaryotic genes (4). At present there are three known classes of DNA sequences at splice sites, found in genes for a chloroplast tRNA (5), yeast tRNAs (6), and vertebrate mRNAs (7–9). Each class has distinctive sequence characteristics. For the splice sites of vertebrate genes, comparisons of a limited set of sequences (7, 8) showed that introns were flanked in every case by 5′-/G-T—(intron)—A-G/-3′. Subsequent sequences have borne this out (9) and have shown that the 5′ ("upstream") and 3′ ("downstream") ends of introns have approximately the following consensus or "optimal" sequences:

5′—(exon)—A-G/G-T-A-A-G-T-A—(intron)
—T-T-T-T-Y-T-T-T-T-T-T-C-T-T-N-C-A-G/G—(exon)-3′.

The upstream and downstream splice sites in nuclear RNA molecules presumably must be brought together for RNA splicing to occur. This could be achieved by intramolecular base pairing, but no appropriate complementary structures have been found reproducibly in the vicinity of splice sites (10). We propose an alternative mechanism in which upstream and downstream splicing sites form base pairs with another RNA molecule which could be a structural component of the splicing enzyme(s). There may be an example comparable to this proposed splicing system in RNase P of *Escherichia coli*, which is a site-specific RNase with an essential RNA component (11).

THE MODEL

We considered that such an effector RNA might be found among the discrete, stable, small nuclear RNAs (snRNAs) that are ubiquitous in eukaryotic cells (12). Sequences have been reported for several of these (13–17). When we compared the sequence of U-1 snRNA from rat hepatoma (13) with sequences of the optimal splice site (Fig. 1), we discovered that the first 10 nucleotides after the U-1 RNA cap are perfectly complementary to the 9 nucleotides of the optimal upstream site and to the tetranucleotide C-A-G/G from the optimal downstream site. The next 11 nucleotides of U-1 are purine-rich and can be aligned with the pyrimidine-rich segment of the downstream site. This striking complementarity with splicing sites is not found in any of the other snRNAs now sequenced, which are U-2 (14), 4.5S (15), and adenovirus VA-I RNAs (16, 17). We therefore propose that U-1 snRNA in the RNA splicing complex forms base pairs with both ends of an intron in a large nuclear RNA molecule as shown in Fig. 2, bringing them into exact register for RNA cutting and splicing.

A number of properties of U-1 snRNA [also called SnD (12)] are consistent with its being a component of the RNA splicing complex. It is present in large ribonucleoprotein particles that are loosely associated with euchromatin (29, 30). U-1 RNA and heterogeneous nuclear RNA (hnRNA) can be co-isolated from these complexes and are dissociated by 70% formamide, which suggests that U-1 RNA is bound to hnRNA by short regions of base pairing (31). U-1 is the most abundant snRNA found in cells whose mRNAs are generated by RNA splicing, including human, mouse, chicken (12, 32), and rat (33) cells. Similar snRNA species also occur in *Xenopus* (34), sea urchin (35), and amoeba (36). U-1 exists in $\approx 10^6$ copies per cell and is as stable as ribosomal RNA (12, 37, 38). Its overall abundance is the same regardless of the transcriptional activity of the cell type (30).

We have fitted all reported sequences of splice sites to the 5′ sequence of the rat U-1 snRNA. At the upstream site, all of the sequences can form at least four base pairs (including some G·U pairs) between U-1 and the beginning of the intron, and at least six such base pairs if the last two nucleotides of the exon are also included. At the downstream site, all the sequences include at least C-A-G/, U-A-G/, or A-G/G, to form base pairs with C-C-U-G in the U-1 RNA. The U-rich, pyrimidine stretch preceding the downstream splice point may also form base pairs with U-1 RNA (e.g., insulin II in Fig. 3), although in most cases an optimal alignment requires looping-out of up to three nucleotides of either the large nuclear RNA or U-1 RNA. [That such looping-out may occur is suggested by the fact that no base is preferred in the fourth position before the splice point (Fig. 1).] Even where the fit of the pyrimidine-rich sequence is poor (e.g., simian virus 40, Fig. 3), there are still three or four contiguous bases pairing with the U-1 sequence C-C-U-G at the downstream splice site. Even a three-base-pair fit should be sufficient to provide precise alignment of the hnRNA, because it is in the codon–anticodon and Shine–Dalgarno (39–41) interactions of mRNA.

There is always at least one nucleotide of terminal redundancy at the ends of an intron, and the alignments illustrated would permit this to contribute to the interaction with U-1 snRNA. That this can happen is suggested by the finding (9) that when an upstream site differs greatly from the optimal sequence /G-T-A-A-G-T . . . in the intron portion it is more likely to conform to . . . A-G/ in the adjacent exon sequence. The terminal redundancy could be used in either of two ways. In the first, branch migration with respect to the standard model of Fig. 2 could allow upstream exon sequences to displace downstream intron sequences from the duplex, or vice versa. Then, the ambiguity in defining the exact splice point from the nucleotide sequence might be carried over into a heterogeneity

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: hnRNA, heterogeneous nuclear RNA; snRNA, small nuclear RNA.

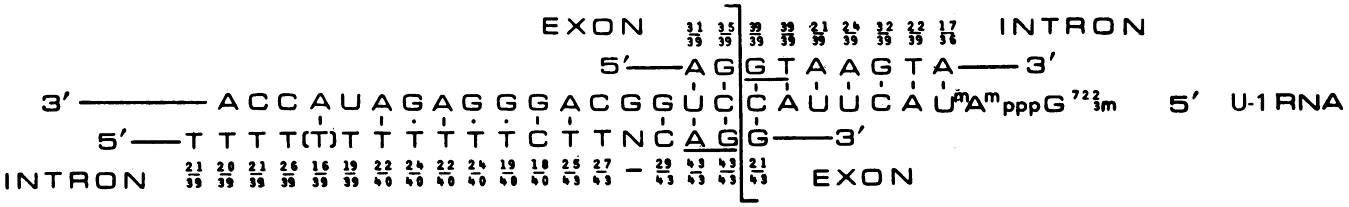
EXON
INTRON
5′——A G G T A A G T A——3′
3′——A C C A U A G A G G G A C G G U C C A U U C A U^{m}A^{m}pppG^{7,2,2}3m 5′ U-1 RNA
5′——T T T T(T)T T T T T T C T T N C A G G——3′
INTRON
EXON

FIG. 1. Consensus or "optimal" upstream (*Top*) and downstream (*Bottom*) splice site sequences, from DNA. These were derived from the data of ref. 9 plus the following extra sequences: human Hb β, γ, and δ (18, 19), rat insulin (20), human growth hormone (21), mouse Ig κ (22–24), and γ2b (25). For intron B of mouse Ig γ1, the sequence from a germ-line clone (26) was used in place of that from the myeloma clone previously reported (27). The latter γ1 gene appears to have suffered a deletion of the first three nucleotides of the intron. Each simian virus 40 and BKV (a human papovavirus) site was entered only once, although they are used in more than one combination *in vivo*. Only one entry was made for each of the following groups: simian virus 40 A4 downstream sites (including deletion variants); the five Ig κ J-region upstream sites; and sites from the small intron in insulin I and II. An intron from silkmoth (28) conforms well to the optimal sites but has not been included because it is so phylogenetically distant from the vertebrate introns. The frequencies of the preferred bases are indicated. Except for one in brackets, all shown are found in ≥45% of sequences. In the pyrimidine-rich stretch leading into the downstream splice site, T is preferred but C is the next most frequent at most positions. Although some introns have longer pyrimidine-rich sequences, the consensus ends 18 nucleotides before the splice site. (*Center*) 5′ sequence of rat U-1 snRNA from ref. 13. G^{7,2,2}3m, N^2,N^2-dimethyl-7-methyl G; A^m, 2′-*O*-methyl A; U^m, 2′-*O*-methyl U. Complementarity to the splice site sequences is indicated.

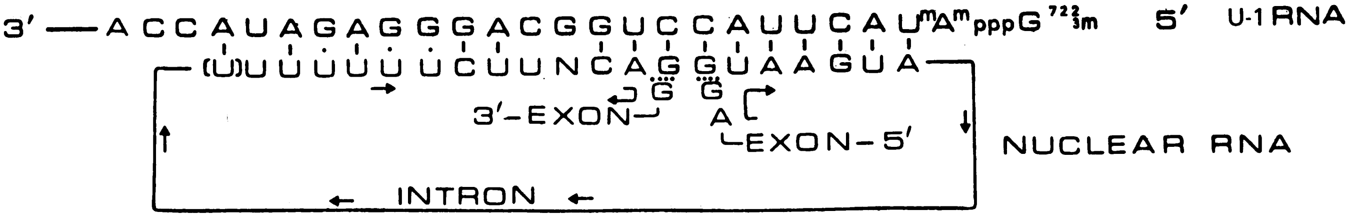
3′——A C C A U A G A G G G A C G G U C C A U U C A U^{m}A^{m}pppG^{7,2,2}3m 5′ U-1 RNA
(U)U U U U U U C U U N C A G G U A A G U A
3′-EXON
EXON-5′
NUCLEAR RNA
INTRON

FIG. 2. Model for base pairing of heterogeneous nuclear RNA to U-1 snRNA in the RNA splicing complex. Dots indicate the predicted points of cutting and splicing.

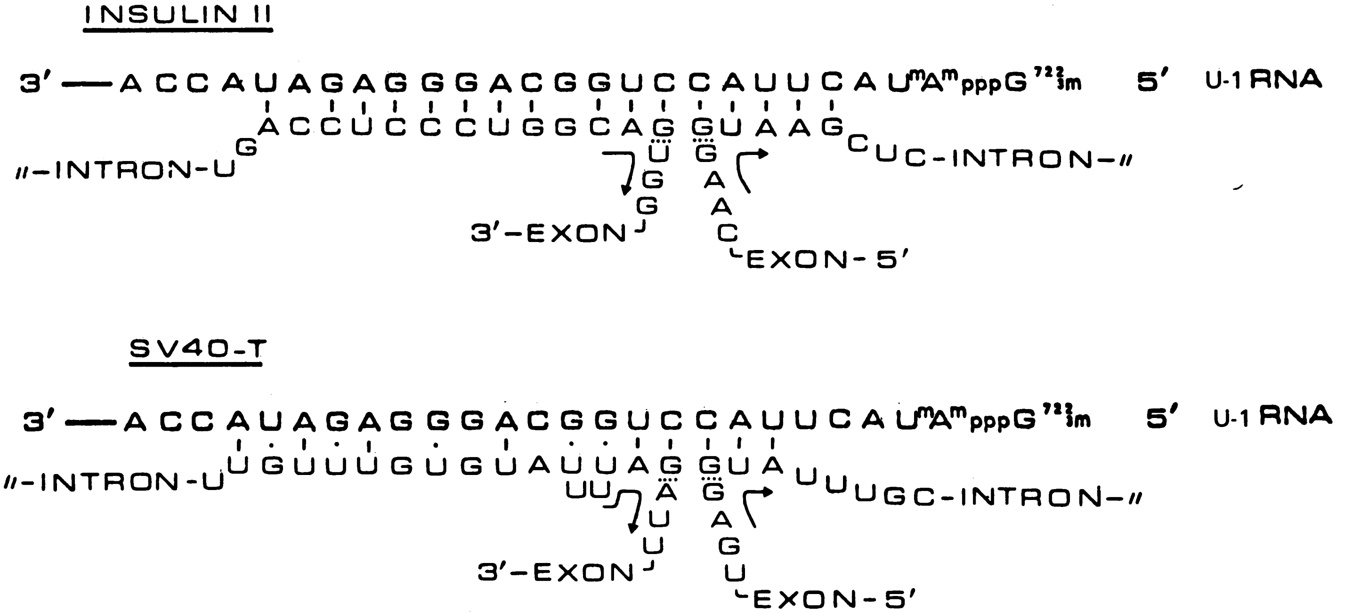
INSULIN II
3′——A C C A U A G A G G G A C G G U C C A U U C A U^{m}A^{m}pppG^{7,2,2}3m 5′ U-1 RNA
A C C U C C C U G G C A G G U A A G
//-INTRON-U G
C U C-INTRON-//
3′-EXON
EXON-5′
SV40-T
3′——A C C A U A G A G G G A C G G U C C A U U C A U^{m}A^{m}pppG^{7,2,2}3m 5′ U-1 RNA
//-INTRON-U U G U U U G U G U A U U A G G U A U U U G C-INTRON-//
3′-EXON
EXON-5′

FIG. 3. Actual sequences fitted to the model. The fit between rat U-1 snRNA and the large intron of insulin II from the same species (20) is one of the best examples. The fit between U-1 and the simian virus 40 (SV40) early region (excision D4-A4 of ref. 9) coding for T-antigen mRNA is one of the poorest examples. In searching for fits, looping-out of either U-1 snRNA or the hnRNA was permitted four nucleotides before the downstream splice point.

in the splice point actually used. In the second, base-pairing of upstream and downstream sites with U-1 could be sequential rather than simultaneous, so that each site would form base pairs maximally, with the splice always being made at the same point.

The preferential splicing of a particular site in a primary transcript containing multiple splice sites could depend on the extent of the base pairing of the intron sequences with U-1 snRNA. Alternatively, the splicing pattern of such a transcript may be determined by other factors acting in concert with U-1 snRNA alignment of splicing sites. The secondary structure of the hnRNA may sequester some potential splice sites and make others more readily available (42, 43). This structure may change as splicing progresses. The geometry of the splicing complex may eliminate other sites by placing a limit on the distance between functionally paired upstream and downstream sites. In this regard, it may be significant that no introns less than 66 nucleotides long have been identified in vertebrate systems. [An exception is a 31-nucleotide excision reported for a minor species of simian virus 40 RNA (44), but the possibility of splicing from a concatameric transcript has not been excluded in this case.] The geometry of the splicing complex may require at least 66 nucleotides of intron for the RNA to fold back on itself as indicated in Fig. 2. Whatever other interactions may be involved, the proposal that U-1 snRNA is the recognition component aligning splicing sites provides a starting point for deciphering the molecular events in RNA splicing.

Note Added in Proof. Lerner *et al.* (45) have also proposed that U-1 snRNA may be involved in RNA splicing.

This research was supported by National Institutes of Health Program Project CA 12800 and Grant AI 13410.

1. Berget, S. M., Moore, C. & Sharp, P. A. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3171–3175.
2. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. (1977) *Cell* **12**, 1–8.
3. Klessig, D. F. (1977) *Cell* **12**, 9–21.
4. Abelson, J. (1979) *Annu. Rev. Biochem.* **48**, 1035–1069.
5. Allet, B. & Rochaix, J.-D. (1979) *Cell* **18**, 55–60.
6. Knapp, G., Ogden, R. C., Peebles, C. L. & Abelson, J. (1979) *Cell* **18**, 37–45.
7. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4853–4857.
8. Catterall, J. F., O'Malley, B. W., Robertson, M. A., Staden, R., Tanaka, Y. & Brownlee, G. G. (1978) *Nature (London)* **275**, 510–513.
9. Seif, I., Khoury, G. & Dhar, R. (1979) *Nucleic Acids Res.* **6**, 3387–3398.
10. Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125–1132.
11. Stark, B. C., Kole, R., Bowman, E. J. & Altman, S. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 3717–3721.
12. Weinberg, R. A. & Penman, S. (1968) *J. Mol. Biol.* **38**, 289–304.
13. Ro-Choi, T. S. & Henning, D. (1977) *J. Biol. Chem.* **252**, 3814–3820.
14. Shibata, H., Ro-Choi, T. S., Reddy, R., Choi, Y. C., Henning, D. & Busch, H. (1975) *J. Biol. Chem.* **250**, 3909–3920.
15. Ro-Choi, T. S., Reddy, R., Henning, D., Takano, T., Taylor, C. W. & Busch, H. (1972) *J. Biol. Chem.* **247**, 3205–3222.
16. Celma, M. L., Pan, J. & Weissman, S. M. (1977) *J. Biol. Chem.* **252**, 9032–9042.
17. Pan, J., Celma, M. L. & Weissman, S. M. (1977) *J. Biol. Chem.* **252**, 9047–9054.
18. Smithies, O., Blechl, A. E., Denniston-Thompson, K., Newell, N., Richards, J. E., Slightom, J. L., Tucker, P. W. & Blattner, F. R. (1978) *Science* **202**, 1284–1289.
19. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
20. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18**, 545–558.
21. Fiddes, J. C., Seeburg, P. H., DeNoto, F. M., Hallewell, R. A., Baxter, J. D. & Goodman, H. M. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4294–4298.
22. Seidman, J. G., Max, E. E. & Leder, P. (1979) *Nature (London)* **280**, 370–375.
23. Max, E. E., Seidman, J. G. & Leder, P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3450–3454.
24. Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. (1979) *Nature (London)* **280**, 288–294.
25. Tucker, P. W., Marcu, K. B., Newell, N., Richards, J. & Blattner, F. R. (1979) *Science*, **206**, 1303–1306.
26. Honjo, T., Obata, M., Yamawaki-Kataoka, Y., Kataoka, T., Kawakami, T., Takahashi, N. & Mano, Y. (1979) *Cell* **18**, 559–568.
27. Sakano, H., Rogers, J. H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. & Tonegawa, S. (1979) *Nature (London)* **277**, 627–633.
28. Tsujimoto, Y. & Suzuki, Y. (1979) *Cell* **18**, 591–600.
29. Zieve, G. & Penman, S. (1976) *Cell* **8**, 19–31.
30. Kano, Y., Komatsu, H., Nakanoin, K. & Fujiwara, Y. (1978) *Exp. Cell. Res.* **115**, 444–447.
31. Flytzanis, C., Alonso, A., Louis, C., Krieg, L. & Sekeris, C. E. (1978) *FEBS Lett.* **96**, 201–205.
32. Marzluff, W. F., White, E. L., Benjamin, R. & Huang, R. C. C. (1975) *Biochemistry* **14**, 3715–3724.
33. Ro-Choi, T. S. & Busch, H. (1974) in *The Cell Nucleus*, ed. Busch, H. (Academic, New York), Vol. 5, pp. 151–208.
34. Rein, A. & Penman, S. (1969) *Biochim. Biophys. Acta* **190**, 1–9.
35. Nijhawan, P. & Marzluff, W. F. (1979) *Biochemistry* **18**, 1353–1362.
36. Goldstein, L. (1976) *Nature (London)* **261**, 519–521.
37. Frederiksen, S., Pedersen, I. R., Hellung-Larsen, P. & Engberg, J. (1974) *Biochim. Biophys. Acta* **340**, 64–76.
38. Weinberg, R. A. & Penman, S. (1969) *Biochim. Biophys. Acta* **190**, 10–29.
39. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
40. Shine, J. & Dalgarno, L. (1975) *Nature (London)* **254**, 34–38.
41. Steitz, J. A. (1979) in *Biological Regulation and Development*, ed. Goldberger, R. F. (Plenum, New York).
42. Rogers, J., Clarke, P. & Salser, W. (1979) *Nucleic Acids Res.* **6**, 3305–3321.
43. Khoury, G., Gruss, P., Dhar, R. & Lai, C-J. (1979) *Cell* **18**, 85–92.
44. Ghosh, P. K., Reddy, V. B., Swinscoe, J., Lebowitz, P. & Weissman, S. M. (1978) *J. Mol. Biol.* **126**, 813–846.
45. Lerner, M., Boyle, J., Mount S., Wolin, S. & Steitz, J. A. (1980) *Nature (London)* **283**, 220–224.