A large inverted repeat sequence overlaps two acceptor splice sites in adenovirus

Stephen H. Munroe

Department of Biology, Marquette University, Milwaukee, WI 53233, USA

Received 15 August 1983; Revised 31 October 1983; Accepted 11 November 1983

ABSTRACT

The distribution of nucleotide sequences resembling functional sites for mRNA splicing was examined by computer-directed searches in order to determine what factors may influence splice site selection in nuclear precursors. In particular, the distribution of large potentially stable hairpin structures or regions of extensive dyad symmetry was studied in adenovirus sequences. One region, spanning 106 nucleotides, was found at 66.4 map units, overlapping back-to-back acceptor sites for two mRNA molecules, those coding for the 100K protein and the 72K DNA binding protein, which are transcribed from opposite strands. This region displays exceptional dyad symmetry and is potentially capable of forming a single, highly stable hairpin when transcribed. It seems likely that the secondary structure as well as the primary structure of RNA plays a role in determining the correct splicing of these mRNA molecules.

INTRODUCTION

Although RNA splicing represents a major event in the biogenesis of many, if not most, mRNA species present in the cell nucleus (1), little is yet known regarding the mechanism of this process. Splicing requires the precise and efficient recognition of sites which mark the 5' and 3' boundaries of exon and intron segments. Cleavage and ligation of adjacent exons take place at these sites. In multiply spliced RNA molecules adjacent, but often widely separated, exons must also be correctly paired. Although certain characteristic sequences are found at the donor (5') and acceptor (3') splice sites of nuclear mRNA precursors (1-3), sequences resembling these canonical donor and acceptor sites occur throughout most transcription units at a frequency nearly equal to that expected for a random sequence (e.g., Table I, Fig. 3 below). Thus canonical donor and acceptor site sequences, at least as presently described, do not provide sufficient information to define unambiguously splice sites within mRNA precursors. Additional information may be provided either by other regions of the primary structure or, alternatively, by the secondary or tertiary folding of the RNA molecule.

In this paper I describe the distribution of both functional splice sites and sites resembling the consensus sequence over a large portion of the adenovirus genome with respect to regions of potentially highly stable intramolecular RNA duplex structure. The possible involvement of an unusually large region of dyad symmetry in the folding and splicing of two major adenovirus mRNA species is discussed.

METHODS

Computer Programs

Primary analysis of published DNA sequences was carried out using 2 programs in the Los Alamos DNA Sequence Analysis Package (4). The first of these, a dyad search routine (5), was used to identify palindromes and inverted repeat sequences (hyphenated palindromes) split by at least 3 and no more than 100 nucleotides. The other program locates potential hairpin structures (6). The maximal overall length of the sequence folded in these runs was ≤ 120 nucleotides. Hairpin loops, internal loops and bulge loops were limited to 20, 5 and 1 nucleotides in length, respectively. Further analysis of selected regions was carried out using the homology search (fh) and alignment (fa) routines (4). The folding of the large hairpin structure located at 66.4 map units was also examined using the RNA5 program developed by Zuker and Stiegler (7).

The distribution of AG and GT dinucleotides conforming to the general requirement of acceptor and donor sites was determined according to the consensus sequences described by Mount (3). Specifically, for acceptor sites AG dinucleotides separated by at least 13 nucleotides upstream from the nearest upstream AG dinucleotide were evaluated as shown in Table I. Donor sites were scored according to Mount (3). A pseudorandom number generator (GGUBS, IMSL Library) was employed to generate a large number of random nucleotide sequences with specified base composition. These were analyzed in parallel to actual sequences.

DNA Sequences

Dyad and secondary structure analyses were carried out on sequences from adenovirus 2 (59.5-62.8, 70.7-100 map units (8-12) and adenovirus 5 (62.8-70.7 map units) (13). Bases are numbered from right to left beginning at the right end (100 map units). Acceptor and donor site analyses also include sequences at the left-hand third of the genome, between 0 and 32 map units (14).

RESULTS AND DISCUSSION**Distribution of Splice Site Consensus Sequences in mRNA Precursors**

The consensus sequences previously described for donor and acceptor sites (1-3) not only fail to provide sufficient information to uniquely determine sites of RNA splicing but actual splice sites, in fact, can represent rather unexceptional fits to these proposed sequences. The pyrimidine composition upstream from 18 known adenovirus acceptor sites (8-15), tabulated in Table I, shows that this group of sites displays a broad range of fits (column 4) based on the variable composition of this characteristic pyrimidine tract. The large number of sites present in both strands of adenovirus which exhibit similar fits to the consensus sequence is shown in column 3. There are 1066 sites which have a pyrimidine composition of 50% or more, over 10 nucleotides. The scoring of a 20 nucleotide tract for pyrimidine provides a more discriminating criterion for locating functional splice sites in the adenovirus genome. Fourteen out of 173 20-nucleotide tracts with $\geq 70\%$ pyrimidine residues have been identified as functional acceptor sites. Despite the pre-

Table I Occurrence in Adenovirus of Sequences Matching the Acceptor Site Consensus Sequence

Pyrimidine Composition Upstream from AG Matching Acceptor Consensus Sequence (1)	Number of Consensus- Sequence Matches		Number of Functional Acceptor Sites Identified (4)
	Random Sequence (2)	Adenovirus Sequence (3)	
A. Pyrimidines within 10 nucleotides			
≤ 4	279 (16)	293	0
5-6	712 (20)	635	6 (1%)
7-8	382 (20)	375	7 (2%)
9-10	31 (6)	56	5 (9%)
B. Pyrimidines within 20 nucleotides			
≤ 10	617 (20)	558	0
11-13	642 (22)	598	4 (1%)
14-16	141 (12)	160	10 (6%)
17-20	4 (2)	13	4 (31%)

Column 1: the number of pyrimidines found within 10 (A) or 20 (B) nucleotides of a possible AG dinucleotide excluding the second nucleotide upstream of the AG which shows no characteristic bias (3).

Columns 2 and 3: in each strand of adenovirus a total of 25,932 nucleotides were searched (72% of the genome). The values in column 2 represent the means (and standard deviations) of 100 similar random sequences.

Column 4: previously characterized functional acceptor sites (8-14) categorized by pyrimidine tract length and composition. The percentage of the consensus sequences at which splicing is found is given in parentheses.

ference for AG preceded by long, exceptionally pyrimidine-rich regions, it is apparent that many more sites match the consensus sequence than are likely to function as splice sites *in vivo*. The number of such sites found in this and other sequences surveyed closely matches that present in randomly generated sequences having the same base composition as can be seen by comparing columns 2 and 3. While it might be expected that the sequences which closely resemble processing sites might interfere with the correct processing of an mRNA precursor, there is no evidence for suppression of such acceptor site-like sequences within transcription units. In fact, tracts containing 9 or 10 pyrimidines out of 10 residues (or 17-19 out of 20) preceded AG residues in the adenovirus genome more frequently than expected by chance. A similar observation also pertains to sequences within adenovirus which show a good fit to the donor consensus sequence (data not shown).

The position of acceptor sites in relation to known promoters and donor sites provides little insight into the nature of the recognition process. Sequences matching the acceptor consensus sequence are found throughout both exon and intron sequences apparently randomly distributed with respect to known functional sites (see Fig. 3 below). Similar observations also apply to the distribution of consensus sequences in other genes (ovalbumin, for example) where overlapping mRNAs and symmetrical transcription probably do not occur. Thus, it appears that splice sites must be recognized by the splicing mechanism on the basis of either additional primary sequence elements, as recently reported for several yeast mRNA species (16), or features of the 3-dimensional folded RNA structure as suggested by variant splicing patterns observed in well characterized mutants (17,18).

Distribution of Potential Hairpin Structures in Adenovirus Transcripts

The location of large inverted repeats in adenovirus DNA as determined by two different search routines is shown in Fig. 1. Both searches focused on large inverted repeats. Since the number of possible intramolecular duplex structures for most RNA sequences is extremely large (16), it is useful to focus on the largest, statistically most significant structures. Although this approach necessarily ignores potentially important short duplex regions, and is to some extent arbitrary, there is evidence that nuclear proteins bound to mRNA precursors destabilize intramolecular duplexes (17). Thus only very stable regions of the RNA secondary structure may exist within RNA-protein complexes in the nucleus.

Fig. 1 shows that the most significant and most stable short-range hairpin structures (B) are distributed throughout the sequence examined. Sig-

Figure 1

Distribution of inverted repeat sequences and potential RNA hairpin structures in right-hand end of adenovirus. The location of inverted repeats (A) and possible RNA hairpins (B) is shown with respect to either the frequency, P , with which these sequences would be expected to occur by chance (A) or the free energy of helix formation (B) as indicated on the ordinate. In both panels the arrows indicate the large inverted repeat structure shown in Fig. 2.

nificance of the structures mapped in Fig. 1A is expressed in terms of a probability function P which is inversely related to the length and symmetry of the dyad (5). Three of the dyad sequences shown in Fig. 1A and 3 of the large hairpins in Fig. 1B overlap one of the 20 known splice sites in this region. This coincidence, however, is close to the level expected for a random distribution since these hairpin structures span a significant fraction (18–26%) of the region examined. Thus, it does not support a general role for RNA secondary structure in marking sites of RNA processing. Taking Fig. 1A and B together, however, reveals one case in which both search routines reveal a structure that overlaps a known splice site. This site at 66.4 map units, marked by arrows in Fig. 1A and B, displays an exceptional degree of symmetry and has a very low free energy for intramolecular helix formation.

Possible Secondary Structure of the 106 Nucleotide Dyad at 66.4 Map Units

The sequence of this large dyad region, which spans 106 nucleotides, is shown in Fig. 2A. The region is centered on not just one, but two, acceptor splice sites arranged back-to-back and located at 66.4 map units. One of these sites, present on the l-strand transcript, represents the 5' end of the body for the early 72K binding protein (13,21,22). The other, on the r-strand transcript, represents the 5' end of the body of the late 100K protein mRNA (13,23). The location of these two mRNA species within a symmetrically transcribed region of adenovirus is shown in Fig. 3. Such close-packing of functionally similar sites may create certain restraints in the primary structure which would be more obvious here than at single, isolated splice sites.

Figure 2

Sequence and possible RNA secondary structure of the large inverted repeat at 66.4 map units. Homology between opposite strands of the 2 halves of the inverted repeat is shown in (A). Boxes indicate homologies between opposing halves, dashes indicate gaps introduced to align homologous segments, dotted lines mark pyrimidine tracts and arrows show position of acceptor sites for mRNA transcripts. Possible hairpin structures formed by l-strand (B) and r-strand (C) transcripts are shown with arrows indicating the splice sites. Initiation codons are underlined.

Fig. 2A gives the sequence of this dyad, illustrating the symmetry about the dyad axis located between bases 11,847/11,848 (shown at right). The two halves of this 106 nucleotide region are homologous at 42 out of 58 positions. Both the AG dinucleotides preceding the 5' end of the bodies of the two mRNAs and 4 unbroken pyrimidine tracts are symmetrically situated about the dyad axis as shown in Fig. 2A. Inasmuch as pyrimidine-rich regions characteristically precede the acceptor splice site within nuclear mRNA species (1,3) to some extent, at least, the symmetry of this region is directly related to the requirements for mRNA splicing. Since a significant portion of this 106 nucleotide region is not included within the pyrimidine tracts some further features of this region may also be important for RNA splicing.

Fig. 2B and C show 2 similar hairpin structures which might be formed by transcripts of the l- and r-strand respectively. Both strands contain 41 out

Figure 3

Summary of mapping data for 72K DNA binding protein and 100K protein mRNAs (13,15,21,22,23). Direction of transcription is indicated by the large arrows. Solid regions indicate exons, open regions introns. Filled circles mark the positions of sequences matching the acceptor site consensus sequence with at least 9/10 or 15/20 pyrimidines preceding the AG. Small arrows mark the positions of functional acceptor sites in these or overlapping mRNA molecules. Only the third leader of the 100K mRNA is shown here.

of a possible 51 base pairs, and a high proportion of GC pairs as well as relatively long stretches of 6-12 uninterrupted base pairs. The overall free energy calculated for helix formation in these hairpins is -70.3 kcal/mole (Fig. 2B) or -72.1 kcal/mole (Fig. 2C) (24-26). Further modeling studies were carried out using RNA5 (7), an efficient RNA folding program which finds the theoretically most stable secondary structure for a given sequence (24). The structure shown in Fig. 2C was also present when sequences of up to 506 nucleotides centered on this dyad were folded. These observations suggest that the RNA secondary structure proposed in Figure 2 may be highly stable under physiological conditions both in terms of the overall free energy of folding and with respect to other possible short-range structures. The perfect 8 base-pair palindrome which spans the 2 splice sites as noted previously (13) accounts for a maximum of only 2 out of 41 base pairs in these structures.

Role of RNA Structure and Dyad Symmetry at Functionally Important Sites

Zain *et al.* (27) have proposed secondary structures for 3 other adenovirus acceptor sites. These model structures include regions adjacent to the 5' ends of the second and third leader sequences of the major late transcription unit and the γ leader of the fiber protein mRNA. In each of these structures, as in those shown in Fig. 2, the acceptor site is found within an unpaired loop region or immediately next to it. Several other recent studies have also suggested that RNA secondary structure adjacent to splice sites is important for splicing. The most detailed of these relate to homologous splice sites within fungal mitochondrial transcripts (28-31) and protozoan rRNA (26). Although the splicing of mitochondrial mRNA and ribosomal transcripts clearly differs in several important respect from that of nuclear en-

coded mRNA, these studies may indicate a general role for secondary structure in RNA splicing.

Secondary structures within mRNA precursors which encompass both intron and exon sequences at the boundary may play one or more possible roles in RNA splicing. Locally stable secondary structure at splice junctions may serve to sequester junction sequences within nuclear RNP complexes in a conformation readily accessible to enzymes, snRNP (32), or RNA binding proteins involved in splicing. Stem and loop structures at these sites may provide specific binding sites for enzymes or complexes. Another, somewhat different, role for secondary structure adjacent to splice sites is that such a structure, possibly in conjunction with RNA-bound proteins, may nucleate the tertiary folding of RNA molecules into a compact structure in which donor and acceptor sites located in different regions of the linear structure are brought closely together.

Finally, it is possible that the extensive dyad symmetry described here is related to some function of this region other than RNA processing. For example, homologies in the amino acid sequence of the N-terminal regions of the 72K and 100K proteins might give rise to a dyad such as this. Inspection of the sequence following the AUG initiation codon, however, shows little evidence of amino acid homology in this region. The most homologous regions of the nucleotide sequence are aligned out of phase due to the introduction of gaps as shown in Fig. 2A. Since the most significant regions of nucleotide homology lie downstream from the AUG codons, it also seems unlikely that translation control regions are involved in the homology of transcripts from both strands in the region. Alignments of the 5' coding sequences of these 72 mRNAs with those of 15 other adenovirus mRNAs have been examined, but no nucleotide sequence homologies were found which were as close as those shown in Fig. 2A. Thus there seems to be no function at the translational level for this extensive dyad sequence within the adenovirus genome. In many organisms and viruses regions of dyad symmetry mark sites important for initiation of DNA replication (33), regulation of transcription (34) or RNA processing (35, 36). Of these possibilities, only the last seems relevant to this particular site.

Both the 72K DNA binding protein mRNA and the 100K protein mRNA represent major products processed from transcription units encoding a relatively large number of overlapping or alternatively spliced mRNA species (reviewed by Flint, ref. 37). Thus both of the homologous acceptor sites at 66.4 map unit sites are frequently utilized in the production of relatively abundant mRNAs

(22,23,38). It seems probable that the secondary structure, as well as the primary structure, of the region bordering these splice junctions plays a role in the recognition and pairing of appropriate exon segments.

ACKNOWLEDGEMENTS

I wish to thank Dr. W. Goad for generously providing access to computer facilities at Los Alamos National Laboratory, Dr. M. Zuker and the National Research Council of Canada for providing RNA folding programs, and Dr. G. Waring for her comments on this manuscript. This research was supported by awards from the N.I.H., the American Cancer Society and Marquette University.

REFERENCES

1. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* 50, 349-383.
2. Sharp, P.A. (1981) *Cell* 23, 643-646.
3. Mount, S.M. (1982) *Nucl. Acids Res.* 10, 459-472.
4. Kanehisa, M. (1982) *Nucl. Acids Res.* 10, 153-162.
5. Goad, W. and Kanehisa, M. (1982) *Nucl. Acids Res.* 10, 247-263.
6. Kanehisa, M. and Goad, W. (1982) *Nucl. Acids Res.* 10, 265-278.
7. Zuker, M. and Stiegler, P. (1981) *Nucl. Acids Res.* 9, 133-148.
8. Akusjarvi, G., Zabielski, J., Perricaudet, M. and Pettersson, U. (1981) *Nucl. Acids Res.* 9, 1-17.
9. Galibert, F., Herissé, J. and Courtois, G. (1979) *Gene* 6, 1-22.
10. Herissé, J., Courtois, G. and Galibert, F. (1980) *Nucl. Acids Res.* 8, 2173-2192.
11. Herissé, J. and Galibert, F. (1981) *Nucl. Acids Res.* 9, 1229-1240.
12. Herissé, J., Rigolet, M., Dupont de Dinechin, S. and Galibert, F. (1981) *Nucl. Acids Res.* 9, 4023-4042.
13. Kruijer, W., van Schaik, F.M.A. and Sussenbach, J.S. (1981) *Nucl. Acids Res.* 9, 4439-4456.
14. Gingeras, T.R., Siakiy, D., Gelinas, R.E., Bing-Dong, J., Yen, C.E., Kelly, M.M., Bullock, P.A., Parsons, B.L., O'Neill and Roberts, R.J. (1982) *J. Biol. Chem.* 257, 13475-13491.
15. Kruijer, W., van Schaik, F.M.A., Speijer, J.G. and Sussenbach, J.S. (1983) *Virology* 128, 140-153.
16. Langford, C.J. and Gallwitz, D. (1983) *Cell* 33, 519-527.
17. Khoury, G., Gruss, P., Dhar, R. and Lai, C.-J. (1979) *Cell* 18, 85-92.
18. Kuhne, T., Wieringa, R., Reiser, J. and Weismann, C. (1983) *EMBO J.* 2, 727-733.
19. Fitch, W.M. (1974) *J. Mol. Evol.* 3, 279-291.
20. Thomas, J.O., Razuiddin, Sobota, A., Boublik, M. and Szer, W. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 2888-2892.
21. Berk, A.J. and Sharp, P.A. (1978) *Cell* 14, 695-711.
22. Chow, L.T., Broker, T.R. and Lewis, J.B. (1979) *J. Mol. Biol.* 134, 265-303.
23. Chow, L.T. and Broker, T.R. (1978) *Cell* 15, 497-510.
24. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) *Nature New Biol.* 246, 40-41.
25. Salser, W. (1977) *Cold Spr. Harb. Symp. Quant. Biol.* 42, 985-1002.
26. Cech, T.R., Tanner, N.K., Tinoco, I., Wier, B.R., Zuker, M. and Perlman, P.S. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 3903-3907.
27. Zain, S., Gingeras, T.R., Bullock, P., Wong, G. and Gelinas, R.C.

- (1979) *J. Mol. Biol.* 135, 413-433.
28. Burke, J.M. and RajBhandary, U.L. (1983) *Cell* 31, 509-520.
 29. Schmelzer, C., Schmidt, C. and Schweyen, R.J. (1982) *Nucl. Acids. Res.* 10, 6797-6808.
 30. Davies, R.W., Waring, R.B., Ray, J.A., Brown, T.A. and Scazzocchio, C. (1982) *Nature* 300, 719-724.
 31. Wollenzien, P.L., Cantor, C.R., Grant, D.M. and Lambowitz, A.M. (1983) *Cell* 32, 397-407.
 32. Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L. and Steitz, J.A. (1980) *Nature* 283, 220-224.
 33. Challberg, M.D. and Kelly, T.J. (1982) *Ann. Rev. Biochem.* 51, 901-934.
 34. Rosenberg, M. and Court, D. (1979) *Ann. Rev. Genet.* 13, 319-353.
 35. Abelson, J. (1979) *Ann. Rev. Biochem.* 48, 1035-1069.
 36. Robertson, H.D. (1982) *Cell* 30, 669-672.
 37. Flint, S.J. (1982) *Biochim. Biophys. Acta* 651, 175-208.
 38. Stillman, B.W., Lewis, J.B., Chow, L.T., Matthews, M.B. and Smart, J.E. (1981) *Cell* 23, 497-508.