MOLECULAR AND CELLULAR BIOLOGY, Aug. 1993, p. 4939–4952
0270-7306/93/084939-14$02.00/0

Vol. 13, No. 8

# Scanning and Competition between AGs Are Involved in 3′ Splice Site Selection in Mammalian Introns

CHRISTOPHER W. J. SMITH,[1,2]* TUNG TAMMY CHU,[1] AND BERNARDO NADAL-GINARD[1]

*Howard Hughes Medical Institute, Laboratory of Molecular and Cellular Cardiology, Department of Cardiology, Children's Hospital, and Departments of Pediatrics and Cellular and Molecular Physiology, Harvard Medical School, Boston, Massachusetts 02115,[1] and Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, United Kingdom[2]**

Received 25 January 1993/Returned for modification 9 March 1993/Accepted 12 May 1993

**In mammalian intron splicing, the mechanism by which the 3′ splice site AG is accurately and efficiently identified has remained unresolved. We have previously proposed that the 3′ splice site in mammalian introns is located by a scanning mechanism for the first AG downstream of the branch point-polypyrimidine tract. We now present experiments that lend further support to this model while identifying conditions under which competition can occur between adjacent AGs. The data show that the 3′ splice site is identified as the first AG downstream from the branch point by a mechanism that has all the characteristics expected of a 5′-to-3′ scanning process that starts from the branch point rather than the pyrimidine tract. Failure to recognize the proximal AG may arise, however, from extreme proximity to the branch point or sequestration within a hairpin. Once an AG has been encountered, the spliceosome can still see a limited stretch of downstream RNA within which an AG more competitive than the proximal one may be selected. Proximity to the branch point is a major determinant of competition, although steric effects render an AG less competitive in close proximity (~12 nucleotides). In addition, the nucleotide preceding the AG has a striking influence upon competition between closely spaced AGs. The order of competitiveness, CAG ≈ UAG > AAG > GAG, is similar to the nucleotide preference at this position in wild-type 3′ splice sites. Thus, 3′ splice site selection displays properties of both a scanning process and competition between AGs based on immediate sequence context. This refined scanning model, incorporating elements of competition, is the simplest interpretation that is consistent with all of the available data.**

Pre-mRNA splicing is a mandatory consequence of the split nature of eukaryotic genes. The production of functional mRNA from primary transcripts requires the accurate removal of the noncoding introns and ligation of adjacent exons. Pre-mRNA splicing provides an important level at which gene expression can be regulated; pairing of different combinations of splice sites leads to the production of tissue-specific mRNAs that code for distinct protein isoforms or that switch off gene expression via the premature termination of open reading frames (52). The splicing pathway proceeds via a well-characterized two-step reaction of successive transesterifications (14, 30, 39). The integrity of open reading frames demands that the sites of these reactions be specified to a high degree. This specificity is achieved by the presence of consensus sequences at the 5′ splice site ([A/C]AG|**GURAGU**) and three elements associated with the 3′ end of the intron, the branch point (YNYURAY), the polypyrimidine tract downstream of the branch point, and the 3′ splice site (**YAG**|G) (invariant nucleotides are shown in boldface; intron boundaries are indicated by vertical bars). These *cis* RNA sequences provide recognition elements for some of the *trans*-splicing factors that form the spliceosome complex within which splicing occurs. *trans*-acting factors include various proteins, as well as the U1, U2, U4/U6, and U5 small nuclear RNA (snRNA)-containing small nuclear ribonucleoproteins (23, 52). The snRNA components are involved in some of the crucial interactions with the consensus recognition elements via RNA-RNA base pairing. U1 base pairs with the 5′ splice site sequence (46, 48, 60, 63), while U2 base pairs with the branch point sequence (31, 60, 61, 64). In yeast cells, U5 RNA has been shown to interact with exon sequences adjacent to both splice sites (28) and U1 has recently been shown to base pair with the 3′ splice site AG (38). While the 5′ splice site, the branch point, and polypyrimidine tract are required early in the splicing pathway and are recognized by established splicing factors, the mode of recognition of the 3′ splice site AG and the stage in the pathway at which recognition occurs remain less well understood.

Early work on the pyrimidine tract and 3′ splice site AG established that although the AG was not absolutely required for the first step in splicing (37, 40), AG mutants were nevertheless often highly impaired in the efficiency of the first step. This suggested that AG recognition could occur before the first transesterification. Moreover, because of the compact arrangement of most introns, in which the branch point and pyrimidine tract both lie within a region of ~18 to 40 nucleotides (nt) upstream of the 3′ splice site AG (14), it was commonly believed that the consensus 3′ splice site, $Y_{10-12}$NYAG, reflected a functionally obligatory juxtaposition of the pyrimidine tract and the AG. That this was not the case became apparent from the instances of some alternatively spliced introns, in which the branch point and associated pyrimidine tract were far upstream (>100 nt) of the 3′ splice site AG (13, 16, 50). In such a configuration, with an extensive pyrimidine tract adjacent to the branch point, in vitro splicing reactions were able to proceed efficiently with a distant downstream AG (17, 36, 50) or even to efficiently undergo the first step in splicing in the complete absence of an AG (53), as in yeast cells (43, 44). Thus, a functional distinction may be drawn, according to the AG requirement

* Corresponding author.

prior to the first step, between AG-dependent and AG-independent introns (36, 38).

Using the AG-independent intron between α-tropomyosin (α-TM) exons 2 and 3, which contains a strong mammalian branch point sequence and an extensive associated pyrimidine tract far upstream of the 3′ splice site (50), we examined the mechanism of 3′ splice site recognition (53). Because of its particular arrangement, this intron allows for the experimental dissection of requirements for the second step in splicing that is not possible with AG-dependent introns. Using a number of mutations of this intron, we found that, as in more conventionally arranged introns, 3′ splice site cleavage and ligation occurred exclusively after the first AG downstream of the branch point, independent of factors such as distance (up to 196 nt tested), the purine-pyrimidine content preceding the first AG, or the nature of the nucleotide immediately preceding the AG. Although some mutations had a quantitative effect on the second step, most notably slower kinetics for 3′ splice site cleavage at a GAG, the qualitative decision to use the first AG was absolute. The effects of several of the more critical mutations were reproduced within a human globin intron. We interpreted these data to suggest that the second step in splicing involved a mechanism in which the spliceosome conducted a linear downstream search (or scan) from the branch point-polypyrimidine tract, until the first AG was encountered, as had previously been suggested (22, 25). Alternative explanations, such as either use of the first AG arising from a higher probability of collision with the spliceosome active site due to restricted diffusion or preference for AGs preceded immediately by an adjacent pyrimidine-rich stretch, seemed less plausible. However, we could not completely rule out these alternative explanations, since many of the alternative AGs used in the mutant transcripts were considerably closer to the branch point than was the wild-type site and/or were preceded by a stronger pyrimidine stretch. Nonetheless, independent support for the scanning model arose from the observation that the insertion of a stable hairpin between the pyrimidine tract and the 3′ splice site AG completely blocked the second step of splicing. The hairpin was over 70 nt from the branch point and 100 nt from the AG and so was unlikely to have been directly interfering with interactions at those sites. Such effects of hairpin structures have also been observed to block translational initiation which is widely believed to occur via cap-to-AUG scanning (20, 21, 34).

We have now extended our investigation of the mechanism for 3′ splice site selection and have tested transcripts that more stringently rule out alternative explanations of our earlier data. In addition, prompted by the examples of various natural introns (2, 4, 5, 11, 55), we have explored the basis of exceptions to the simple rule that the first AG downstream of the branch point is selected exclusively as the 3′ splice site. We have found that although such exceptions do exist, they are predictable and can be accommodated by simple refinements of the model. This is reminiscent of the scanning model for translational initiation, which has also been able to accommodate exceptions with simple rules to break the rule (21). In the light of the results reported here, a refined scanning model can be summarized thus: scanning initiates at the branch point and proceeds in a 3′ direction to the first AG unless it either is so close to the branch point that recognition is very inefficient or is hidden within stem-loop structure. In either of these cases the first AG can be bypassed and the next downstream AG can be located (as long as the stem-loop is not in a position where it blocks the second step in splicing). Once an AG has been recognized, however, the spliceosome can still "see" a limited stretch of downstream RNA. Within this stretch of RNA the most competitive AG is selected as the 3′ splice site. Determinants of competitiveness include proximity to the branch point, the more proximal AG usually being more competitive, unless it is very close to the branch, in which case steric effects can render it less competitive. The nucleotide preceding the AG also has a marked effect upon competition. The hierarchy of competitiveness, CAG ≃ UAG > AAG > GAG, closely follows the occupancy of this position in consensus compilations of 3′ splice sites (25, 29, 47). This reformulation of the scanning model, incorporating elements of competition, is the simplest interpretation that is consistent with all of the data.

## MATERIALS AND METHODS

**Constructs.** Constructs used as templates for in vitro transcription were all prepared by standard cloning procedures. The parental construct, pGC+DX, contains α-TM exons 2 and 3 with a 28-nt spacer just downstream of the 5′ splice site to relieve steric interference (50, 53). Constructs pR2, ΔPy9, and 23HP have been described previously (53). ΔPy9R2 contained the 5′ half of ΔPy9 with the truncated pyrimidine tract, fused to the 3′ half of pR2 with the purine-rich stretch before the 3′ splice site. pRnAG had an AAG preceded by a purine-rich tract inserted as oligonucleotides into the *Acc*I site in the middle of the intron in pGC+DX. 23HP was derived from pGC+DX by insertion of the hairpin-forming sequence $GGG(CGG)_6GAATTC(CCG)_6$ CCC into the *Acc*I site. 23HPA and 23HPB had the 5′ or 3′ half of the hairpin of 23HP only. They were derived from 23HP by deleting the sequence from the *Eco*RI site in the middle of the hairpin to the *Acc*I site. 23L had the additional loop-forming sequence GAATTCCCCAGCTGCCCAATTC inserted into the *Eco*RI site of 23HP. 23L5′ had the 3′ half of the stem loop of 23L deleted between the *Pvu*II and *Acc*I sites. 23HPBA was derived from 23HPB by insertion of the sequence $GGG(CGG)_6GAATT$ between a blunted *Xba*I site and the *Eco*RI polylinker site downstream of exon 3. 23HPAA was cloned by inserting the same fragment into 23HPA. Constructs ΔPy39HP to ΔPy1HP had the stem-loop-forming sequence $GGG(CGG)_6GAATTC(CCG)_6CCC$ immediately downstream of the 3′-end-truncated pyrimidine tracts of ΔPy39 to ΔPy1 (53). Δ9S was derived from ΔPy9 by insertion of the AG-containing hairpin $GGG(CGG)_6GAATT$ $CAGCCCCCGGGGGCTGAATTC(CCG)_6CCC$ into the *Acc*I site. In Δ9S5′, the 3′ half of the hairpin was removed, leaving the insert $GGG(CGG)_6GAATTCAGCCCCC$. Constructs −12NAG, −6NAG, −3NAG, and +7NAG were cloned by polymerase chain reaction-directed mutagenesis (53). Oligonucleotides degenerate at the appropriate position with respect to the 3′ splice site were used to amplify short segments of DNA which were then recloned into the parental construct. Individual clones were identified by dideoxy sequencing, and the sequence of the entire amplified insert was also verified. −20GAG, −30GAG, and −75GAG were also prepared by polymerase chain reaction-directed mutagenesis, but without degenerate oligonucleotides. −108GAG was prepared by inserting a *Sac*I linker, CGAGCTCG, into the *Acc*I site of pGC+DX (53). pΔ30-110 was cloned by cutting pGC+DX with *Xho*I and *Acc*I and reinserting oligonucleotides bearing the sequence TCGAGGGCGCGCGGT GTGGCACTGCACGT. This resulted in the deletion of the branch point and polypyrimidine tract, leaving a unique *Acc*I site into which oligonucleotides could be inserted. Oligonu-

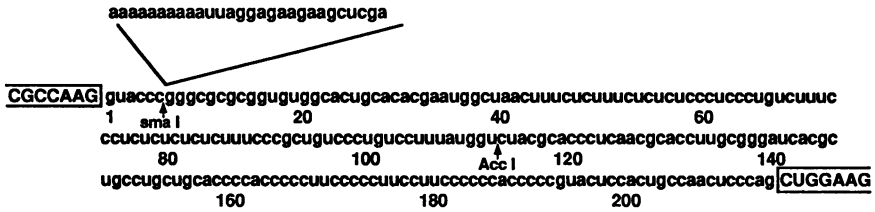



FIG. 1. Sequence of α-TM intron. The sequence of the intron between α-TM exons 2 and 3 is shown, along with the 28-nt spacer inserted into the *Sma*I site just downstream of the 5′ splice site in construct pGC+DX. Features to note are the branch point sequence GGCUAAC at nt 36 to 42 and the polypyrimidine tract at nt 43 to 90.

cleotide inserts contained the 5′-end sequence CTGC in addition to the branch point and downstream sequences GGCUAAC . . . shown in Fig. 7A.

**In vitro transcription and splicing reactions.** $^{32}$P-labelled RNA transcripts were transcribed from pGEM vectors with Pharmacia SP6 polymerase by using a G(5′)ppp(5′)G dinucleotide primer as described previously (1, 50). HeLa cell nuclear extracts were prepared by using the modifications of Abmayr et al. (1). Standard splicing reaction mixtures contained 20 to 50 fmol of [$^{32}$P]RNA transcript, 3.2 mM $MgCl_2$, 500 mM ATP, 20 mM creatine phosphate, 1,000 U of RNasin per ml, 12.8 mM *N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid (HEPES) (pH 8), 14% (vol/vol) glycerol, 62 mM KCl, 0.12 mM EDTA, 1.7 mM dithiothreitol, and 60% nuclear extract and were incubated for 3 h at 30°C. Departures from these standard conditions are noted in the figure legends. Splicing reaction mixtures were then subjected to proteinase K digestion and phenol-chloroform extraction. Further characterization of reaction intermediates and products was carried out by debranching in the HeLa cell S-100 cytoplasmic fraction at 30°C for 30 min. Reaction products were analyzed by electrophoresis in 8 M urea–polyacrylamide gels followed by autoradiography. For each series of transcripts, the sizes of all RNA species were determined both prior to and after debranching. The sizes of all bands corresponded to those expected from use of the predicted splice sites, with the exception of some of the RNA species that had extensive intramolecular secondary structure. Even in these latter cases, the linear bands without secondary structure had the expected mobilities, while lariats could be identified by a change in electrophoretic mobility after debranching, even though the debranched bands still had anomalous mobility. Except for minor species, noted below, there was no evidence for use of cryptic splice sites.

## RESULTS

In the present study, we have used exons 2 and 3 of the rat α-TM gene with a spacer element in the intron to relieve steric interference between the 5′ splice site and branch point (Fig. 1). The branch point in the α-TM intron is 172 nt upstream of the 3′ splice site AG, which is not required before step 2 in splicing (53). In most of these experiments a single AG was used exclusively as the 3′ splice site. We have therefore tended to concentrate upon the identity of the AG selected rather than upon the rate at which the second step occurs. The reason for this emphasis is that although both qualitative and quantitative effects may be instructive about the biochemistry of splicing, the qualitative effects (i.e., the identity of the splice site[s] selected) have a greater impact upon gene expression, given that splicing is not usually rate limiting for mRNA accumulation (reference 35 and our unpublished observations).

**Exclusive use of the first AG does not require adjacent pyrimidines.** We previously demonstrated exclusive use of the first AG downstream of the branch point-pyrimidine tract in the second step of splicing (53). However, in that report all of the AGs upstream of the wild-type 3′ splice site were preceded by regions that were more pyrimidine rich than the area adjacent to the wild-type site. To test whether these AGs were selected because of the adjacent pyrimidine-rich sequences or their greater proximity to the branch point-polypyrimidine tract, we made a number of transcripts in which the first AG was preceded by purine-rich stretches (Fig. 2). We previously observed that changing the content of the 24 nt preceding the 3′ splice site of the α-TM intron 2 from predominantly pyrimidine (70%) to predominantly purine (70%) had no significant effect upon the second step in splicing (53) (Fig. 2, transcripts GC+DX and R2). It is possible that the presence of the very strong upstream pyrimidine tract overrides the effects of purine substitutions adjacent to the AG (36). We therefore made the equivalent purine substitution in a construct in which the functional pyrimidine tract adjacent to the branch point has been decreased from 50 to 9 consecutive pyrimidines (53). Once again, we observed no significant difference in the second step in splicing upon purine substitution adjacent to the 3′ splice site, even though the pyrimidine tract adjacent to the branch point has been truncated to the point of marginal efficiency (Fig. 2, transcripts ΔPy9 and ΔPy9R2). Thus, use of a CAG widely separated from the branch point does not require the AG to be preceded by a pyrimidine-rich stretch, even when the upstream branch point-associated pyrimidine stretch is relatively weak.

We then tested whether AGs inserted upstream of the wild-type site need to be preceded by a pyrimidine-rich stretch in order to be used exclusively. In transcript RnAG, an AG, preceded by a purine-rich stretch, was inserted ~100 nt upstream of the wild-type site. The purine-rich region and the immediate sequence context of this AG (AAGU) both diverge radically from the consensus for a 3′ splice site. Nevertheless, in this transcript we saw exclusive use of the proximal AG and no use of the downstream wild-type site, which has a more favorable match to consensus sequences. That the first AG downstream of the branch point can be used as the 3′ splice site even when preceded by a purine-rich region demonstrates the importance of relative proximity to the branch point-pyrimidine tract as a major determinant in 3′ splice site selection. The relatively slow kinetics of the second step for this mutant, despite the greater proximity of the branch point to the inserted AG than to the wild-type 3′ splice site, is most likely due to the A immediately preceding the proximal AG (see below) and the general purine richness of the preceding region.

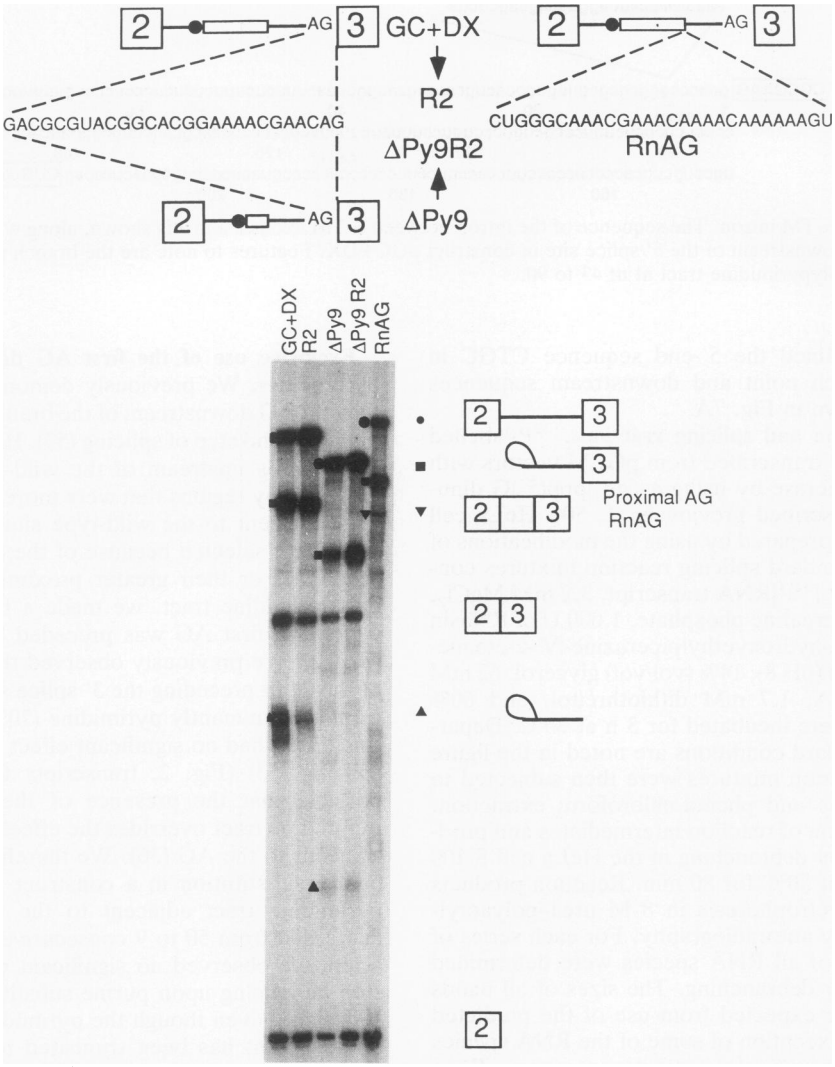



FIG. 2. Exclusive use of the first AG does not require adjacent pyrimidines. Transcript ΔPy9 was derived by deletion of nt 56 to 110 from pGC+DX (Fig. 1), resulting in truncation of the polypyrimidine tract to 9 nt (53). As indicated by the arrows, transcripts R2 and ΔPy9R2 were derived from GC+DX and ΔPy9, respectively, by substitution of the 25 nt immediately preceding the 3′ splice site AG to the purine-rich sequence shown. This results in a change in the pyrimidine content in this region from 70 to 30%. Both R2 and ΔPy9R2 undergo the two steps of splicing with efficiency and specificity comparable to those of the parental transcripts. In transcript RnAG, an AG preceded by a purine-rich tract is inserted in the *Acc*I site at nt 110 within the intron (Fig. 1). Use of the downstream wild-type AG is occluded, and the proximal AG is used exclusively, yielding a spliced product of 404 nt, as indicated by the triangle. Splicing reaction mixtures were incubated for 3 h and run on a 5% polyacrylamide gel.

**Evidence for a scanning process in 3′ splice site selection.** The preceding data demonstrate that an adjacent pyrimidine-rich region is not essential for exclusive use of the first AG. Nevertheless, the exclusive selection of a proximal AG need not arise as the result of a linear scan but rather arises from a constrained diffusion-controlled collision, in which the RNA between the branch point and AG is looped out. A more proximal AG would be favored because of the greater probability of collision with its recognition site within the spliceosome. To distinguish between these possibilities, constructs with very closely spaced CAGs, over 170 nt from the branch point but separated by only 6 or 7 nt, were monitored for their selection as a 3′ splice site (Fig. 3A, −6CAG and +7CAG). The constrained-diffusion model predicts nearly equal use of the competing AGs, whereas a simple scanning model predicts predominant use of the first AG. Both transcripts showed exclusive use of the first AG (Fig. 3A, +7CAG and −6CAG), providing strong support for a scanning model over a constrained-diffusion model.

We next produced a number of constructs in which secondary structure was introduced between the branch point-pyrimidine tract and the 3′ splice site AG. We previously showed that a stable hairpin between the branch point and 3′ splice site completely inhibited the second step in splicing both in vitro and in vivo (Fig. 3B, 23HP) (49, 53). Control transcripts containing only the 5′ or 3′ side of the stem-forming sequences, in contrast, spliced perfectly well using the downstream AG (Fig. 3B, 23HPA and 23HPB, respectively). This result is again consistent with a scanning model but not a constrained-diffusion model. Next, we produced a transcript in which a proximal CAG was placed within a loop at the end of the hairpin of 23HP (Fig. 3B,

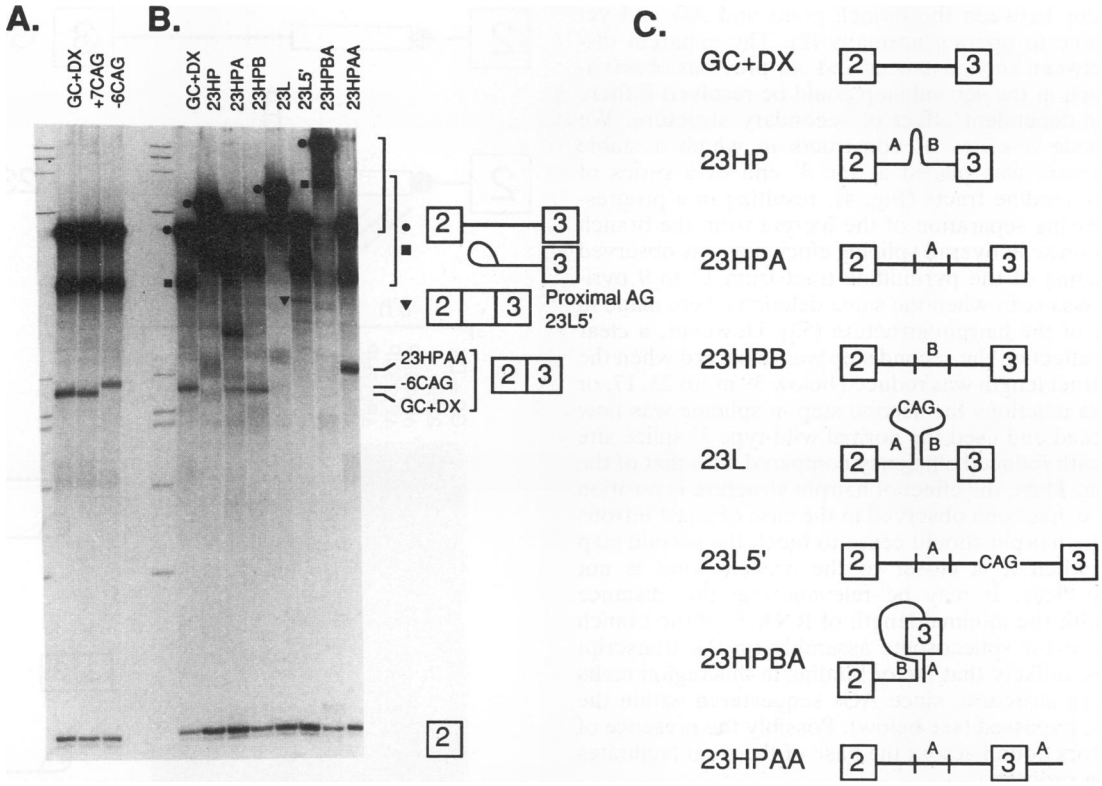



FIG. 3. Evidence for a scanning process in the second step in splicing. (A) Transcripts −6CAG and +7CAG were derived from pGC+DX by mutagenesis and contain CAGs 6 nt upstream and 7 nt downstream of the wild-type CAG, respectively. In each case only the first AG is used, leading to spliced products of 307 nt (−6CAG) and 301 nt (GC+DX and +7CAG). Use of the second AG in +7CAG would have yielded a product of 294 nt. Splicing reaction mixtures were incubated for 4 h and run on a 5% polyacrylamide gel. (B) The transcripts shown schematically in panel C were spliced for 4 h and run on a 5% polyacrylamide gel. 23HP contained a stable hairpin inserted at the *Acc*I site downstream of the pyrimidine tract (Fig. 1), which blocked the second step of splicing. 23HPA and 23HPB had only the 5′ or 3′ half of the stem-forming sequences and were able to splice to the wild-type AG. The band of ~350 nt in the 23HPA lane was present in the unprocessed transcript and is not related to splicing. Transcript 23L had the loop of 23HP extended with a CAG in the middle of a 12-nt loop. The CAG was not used as a 3′ splice site; the second step of splicing was blocked. When the 3′ half of the stem loop of 23L was deleted in 23L5′, the wild-type site was still not used, but a new spliced product of 404 nt (marked with a triangle) was observed, indicative of use of the proximal CAG. Transcript 23HPBA had stem-forming elements in the middle of the intron and on the 3′ side of exon 3. The 3′ half of the intron and exon 3 would therefore be sequestered in a large (>200-nt) loop as diagrammed in panel C. This transcript showed no second step in splicing, consistent with blockage of scanning by the stem structure. Control experiments established that intermolecular base pairing between stem-forming sequence elements, which could account for the apparent lack of spliced product with 23HPBA, does not occur under the gel electrophoresis conditions used. A control transcript, 23HPAA, had direct repeats in the same locations and efficiently underwent the second step of splicing at the wild-type CAG, yielding a spliced product of 325 nt. (C) Schematic representation of transcripts in panel B. The sequence elements forming the 5′ and 3′ halves of the hairpin of 23HP have been labelled A and B, respectively.

23L). If proximal AGs are favored because of constrained diffusion, we would expect use of the proximal loop AG as the 3′ splice site, whereas a scanning model predicts that neither AG would be used. Transcript 23L showed a block in the second step in splicing, with spliced product produced from neither AG. In contrast, a control transcript, 23L5′, in which the 3′ half of the stem-forming structure was deleted, showed exclusive use of the proximal AG. In transcript 23L the proximal AG was in a loop of only 12 nt. It is possible that the small size of the loop prevented use of the proximal AG. However, we obtained a similar result using a transcript in which the wild-type 3′ splice site CAG was sequestered in a loop of >200 nt. In this transcript (Fig. 3B, 23HPBA), one of the inverted repeats that form the stem was in the same position between the pyrimidine tract and 3′ splice site, while the other was at the 3′ end of α-TM exon 3. This transcript also showed no production of spliced products, again consistent with scanning but not with alternative models for 3′ splice site selection. A control transcript with direct repeats, unable to form the stem-loop, spliced efficiently using the wild-type site (Fig. 3B, 23HPAA).

The preceding data showed that the second step in pre-mRNA splicing has characteristics that are more consistent with a scanning mechanism than with alternative explanations. However, a number of cases in which the first AG downstream of the branch point-pyrimidine is bypassed, either partially or completely, for use of a more distal AG have been documented (2, 5, 8, 36, 58). We therefore proceeded to investigate whether such exceptions are inconsistent with or can be accommodated with a scanning-like process.

**Position-dependent effects of a stable hairpin structure on the second step in splicing.** We have shown that stable hairpin-forming sequences between the polypyrimidine tract and 3′ splice site AG can lead to a block in the second step in splicing (Fig. 3) (53). However, in some introns, natural

hairpins occur between the branch point and AG and yet splicing is able to operate normally (2). The apparent discrepancy between such instances and our previous observation of a block in the second step could be resolved if there is a position-dependent effect of secondary structure. We therefore made a series of constructs in which a stable hairpin structure was placed at the 3′ end of a series of truncated pyrimidine tracts (Fig. 4), resulting in a progressively decreasing separation of the hairpin from the branch point. A decrease in overall splicing efficiency was observed upon shortening of the pyrimidine tract from 17 to 9 pyrimidines, as was seen when the same deletions were made in the absence of the hairpin structure (53). However, a clear and distinct effect on the second step was observed when the pyrimidine tract length was reduced below 39 nt, to 23, 17, or 9 nt; in these deletions the second step in splicing was now able to proceed and used the normal wild-type 3′ splice site AG, albeit with reduced efficiency compared with that of the native intron. Thus, the effect of hairpin structure is position dependent, as has been observed in the case of yeast introns (15). Why the hairpin should cease to block the second step of splicing when it is closer to the branch point is not immediately clear. It may be relevant that this distance correlates with the minimal length of RNA 3′ of the branch point required for spliceosome assembly on the transcript (44, 53). It is unlikely that factor binding in this region melts the secondary structure, since AGs sequestered within the structure are bypassed (see below). Possibly the presence of splicing factors bound across the base of the stem facilitates the scanning process.

**Bypassing of proximal AGs sequestered in hairpin structures.** The distance within which we have determined that hairpin structures do not block the second step in splicing correlates with the known examples of yeast and viral introns that have such structures between the branch point and 3′ splice site (2, 5). To confirm that AGs sequestered in hairpins in such a location can indeed be bypassed in favor of downstream AGs, we constructed mutant pΔ9S (Fig. 5), in which the stem downstream of the 9-nt pyrimidine tract contained a CAG sequence motif. In vitro splicing of this transcript showed complete bypassing of the proximal AG and use of the wild-type distal 3′ splice site AG. In a control construct, pΔ9S5′, the 3′ half of the hairpin was deleted, leaving the proximal AG intact. In this transcript, use of the distal AG was undetectable and the proximal AG was now used exclusively, albeit with slow kinetics. Thus, when a hairpin is in a position downstream of the branch point in which it does not block the second step, AGs sequestered within the structure are bypassed and the next downstream AG is selected. Similar results were obtained when the CAG was within the loop of the hairpin structure (data not shown). These data therefore confirm that proximal AGs can be bypassed and more-distal AGs can be selected when the proximal AG is sequestered within a secondary structure, as has been seen in both viral (2) and yeast (5) introns. It should be emphasized, however, that although secondary structures close to the branch point can be bypassed, along with any sequestered AGs, the striking behavior of hairpins distant from the branch point and AG is still strongly supportive of a 5′-to-3′ scanning process.

**Competition between closely spaced AGs based on the preceding nucleotide and relative position.** In a number of experimental test constructs and native introns it has been observed that more than one of a series of closely spaced AGs may be used as a 3′ splice site (8, 36, 58). Indeed, in some cases such use of alternative in-frame AGs can give rise to the insertion of additional codons into the reading frame of the mRNA (e.g., see references 4, 11, and 55). In the experimental transcripts (8, 36, 58) the upstream AG was preceded by an A or G, and it has been suggested that the presence of the nonconsensus purine at this position allowed use of both AGs rather than exclusively the proximal AG (36). In order to systematically investigate the roles of distance of separation and of the preceding nucleotide in the ability of closely spaced AGs to compete, we made the series

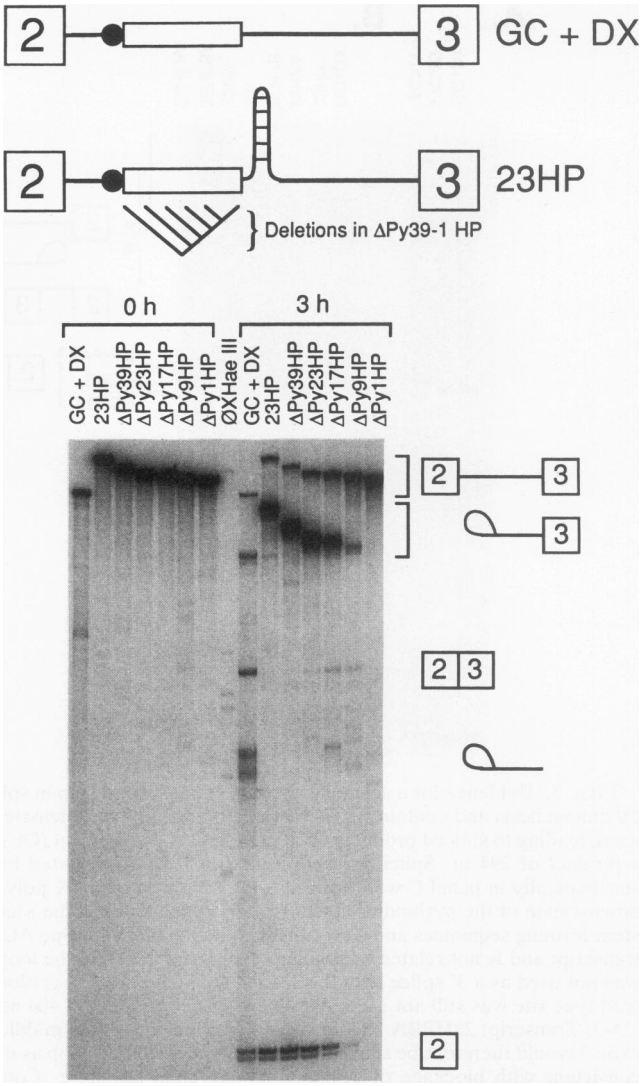



FIG. 4. Position-dependent effects of hairpin structure upon the second step in splicing. Stable hairpin-forming sequences were inserted at the 3′ ends of various truncated pyrimidine tracts. 23HP was derived by the insertion of hairpin-forming sequences into the *Acc*I site of GC+DX (see Fig. 1 and Materials and Methods). The series of transcripts ΔPy39HP to ΔPy1HP had the same hairpin sequences inserted at the ends of pyrimidine tracts progressively truncated to between 39 and 1 nt. Transcript RNA was incubated under standard splicing conditions for 0 or 3 h and run on a 5% polyacrylamide gel. Scanning is blocked when the hairpin is at the 3′ end of the full-length or 39-nt pyrimidine tract, but spliced product at 301 nt can be seen when the pyrimidine tract has been shortened to 23, 17, or 9 nt. ΔPy1HP shows no activity for either step in splicing because of the complete deletion of the pyrimidine tract.

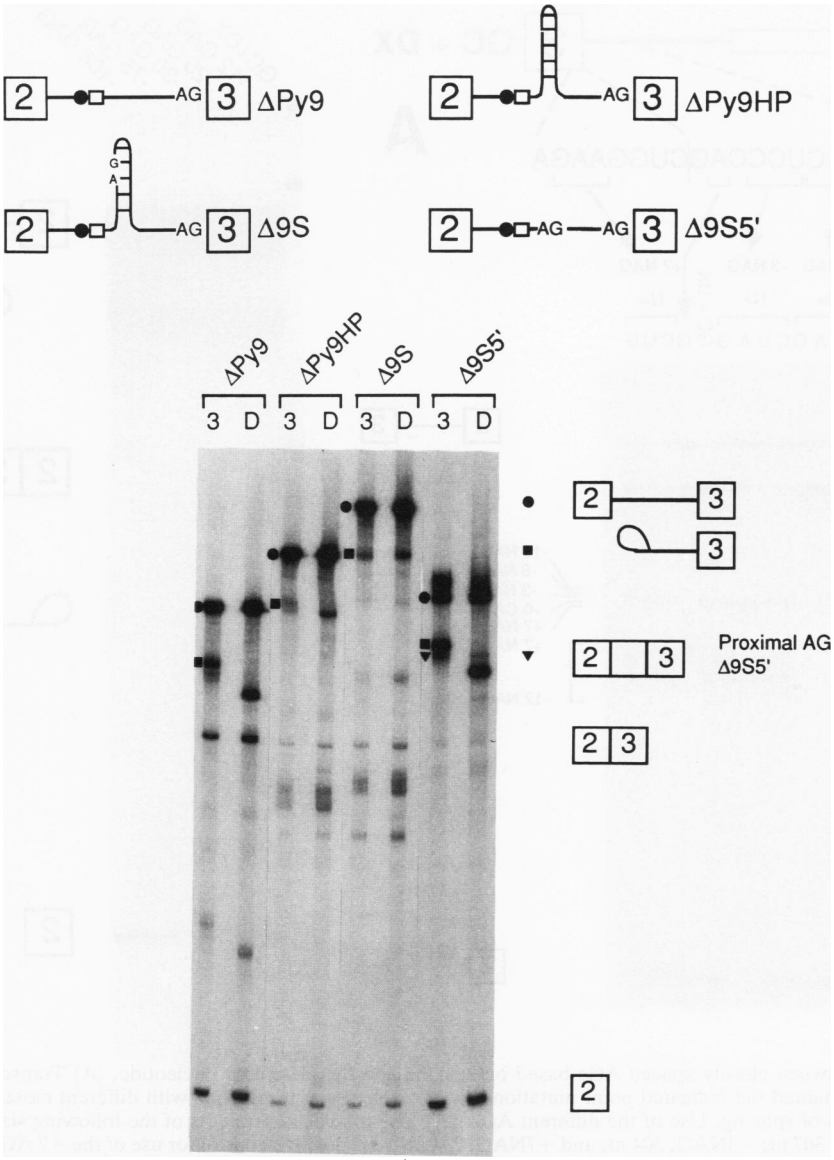



FIG. 5. Bypassing of an AG sequestered in a hairpin structure. Transcripts are shown schematically at the top of the figure and are described in more detail in Materials and Methods. Briefly, Δ9S was derived from ΔPy9HP by the insertion of additional hairpin-forming sequences containing a CAG on the 5′ side of the stem. Δ9S5′ had the 3′ half of the hairpin deleted. Splicing was carried out for 3 h under standard conditions except that $MgCl_2$ was at 2 mM. Debranched reactions are shown in lanes D. Transcript Δ9S5′ used the proximal AG to give a spliced product of 407 nt, marked by the triangle. The proximal AG in Δ9S would have given a spliced product of 437 nt. However, this AG was bypassed and the distal AG was used, yielding the 301-nt spliced RNA. In lanes D, the lariats produced by ΔPy9 and Δ9S5′ run with increased mobility compared with that of the undebranched samples (lanes 3) (branched lariat intermediates are marked by squares), as expected. However, the extensive hairpin in ΔPy9HP and Δ9S transcripts and corresponding lariats results in anomalously low mobility, even after debranching.

of constructs −12NAG, −6NAG, −3NAG, and +7NAG (Fig. 6A). In these transcripts, competing AGs preceded by each of the four nucleotides C, U, A, and G were placed 12, 6, or 3 nt upstream of the wild-type 3′ splice site AG or 7 nt into the exon, respectively. The results of in vitro splicing reactions are shown in Fig. 6A. Once again, the use of the various AGs can easily be monitored by the variation in size of the spliced product. A number of conclusions can be made from these data. First, there is a clear effect of distance between competing AGs; in general, the greater the separation is, the greater is the likelihood that the upstream AG will be used, consistent with scanning. At the −3 position, CAG and UAG are used in about a third of transcripts with a slight preference for the downstream CAG, while in the −6 or −12 position they are used almost exclusively. On the other hand, while AAG is bypassed at the −3 position, it competes well at −6 and −12. Second, the identity of the preceding nucleotide has a clear influence on both the competitiveness of AGs and the kinetics of the second step. Thus, at the −3 position, CAG and UAG compete well, while neither AAG nor GAG is used at all. At −6, where CAG and UAG are used exclusively, AAG does compete although some tran-

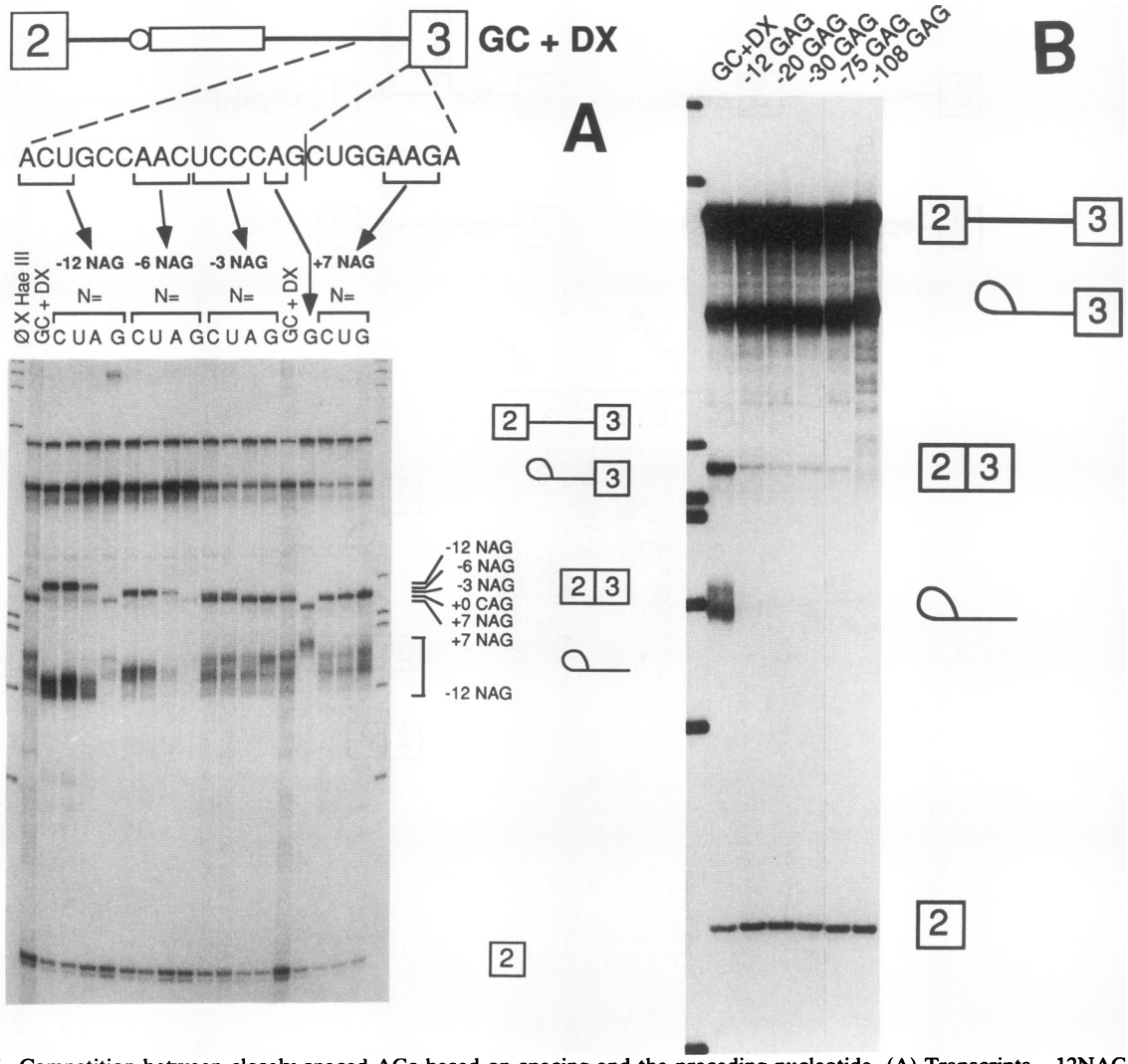
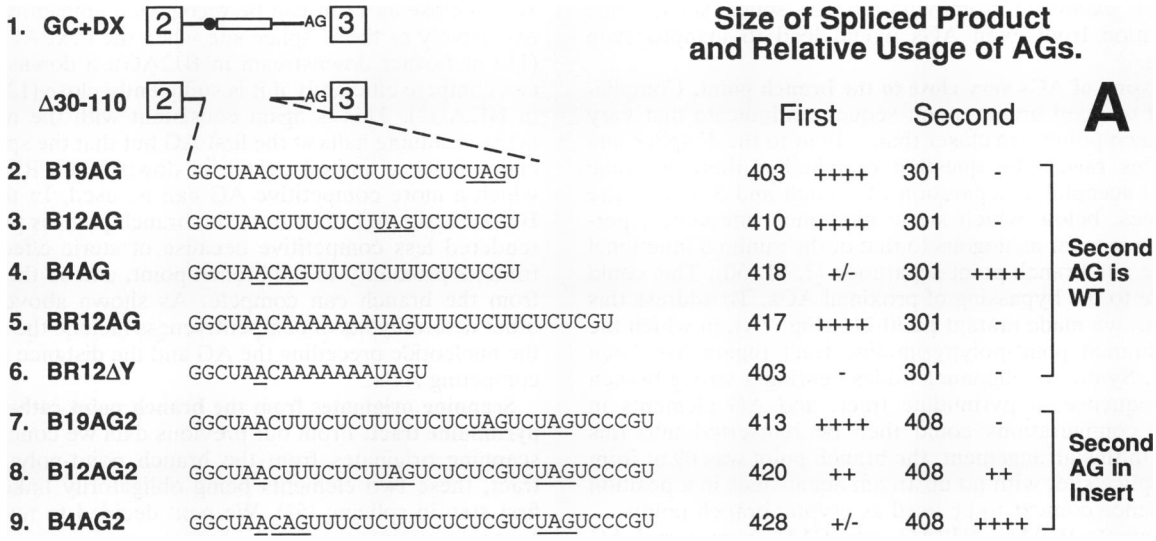



FIG. 6. Competition between closely spaced AGs based on spacing and the preceding nucleotide. (A) Transcripts −12NAG, −6NAG, −3NAG, and +7NAG contained the indicated point mutations to give a series of transcripts with different closely spaced AGs that could compete in the second step of splicing. Use of the different AGs gave rise to spliced products of the following sizes: wild-type AG, 301 nt; −12NAG, 313 nt; −6NAG, 307 nt; −3NAG, 304 nt; and +7NAG, 294 nt. As a positive control for use of the +7 AG, lane G contained a point mutation converting the wild-type splice site AG to GG, resulting in efficient use of +7AAG as the 3′ splice site. The lariat products also varied in size in accordance with the AG used and are indicated by the bracket at the side of the gel. However, the size variation is not as clear as with the spliced products, because of heterogeneity of the lariat products, presumably arising from exonuclease nibbling of the lariat tails. In the −12GAG lane, a small amount of a lariat intermediate can be seen above the transcript. This results from the activation in the −12NAG series of a cryptic branch point by bringing the AG close to the 3′ end of the highly pyrimidine-rich stretch at nt 172 to 191 of the intron (Fig. 1). This is seen only in the −12GAG lane because of low rate of the second step, which leads to accumulation of intermediates. It should be noted that the low-level activation of this cryptic branch point has no impact upon conclusions about second-step scanning. (B) The indicated transcripts had GAGs at the designated positions upstream from the wild-type CAG. Use of the wild-type CAG is reduced in all mutant transcripts but becomes undetectable only in the −108GAG lane.

scripts still use the wild-type 3′ splice site. Finally, in the +7 series none of the different NAGs is able to compete with the wild-type CAG located 7 nt upstream, even though the AAG at this position is used efficiently when the wild-type 3′ splice site CAG is mutated to CGG (Fig. 6A, lane G). Assessing the degree to which AAG and GAG compete is rendered problematic by the clear effect that these nucleotides also have on the kinetics of the second step in splicing. At this position G, and to a lesser extent A, has previously been shown to slow down the second step in splicing (53). In the present experiments, this kinetic effect has the consequence that in some transcripts the competitiveness of a GAG or AAG will be underestimated by examination of spliced product at a single time point. Thus, in the −6GAG and −12GAG lanes the lariat intermediate can be seen to be accumulating, consistent with selection of the GAG, even though cleavage and ligation at this site are too slow to be detectable, at least after 3 h. These caveats notwithstanding, it is possible to formulate a hierarchy of competitiveness of the different NAGs: CAG ≃ UAG > AAG > GAG. This hierarchy parallels the consensus preference in natural introns, with the exception that wild-type 3′ splice sites also show a clear preference for



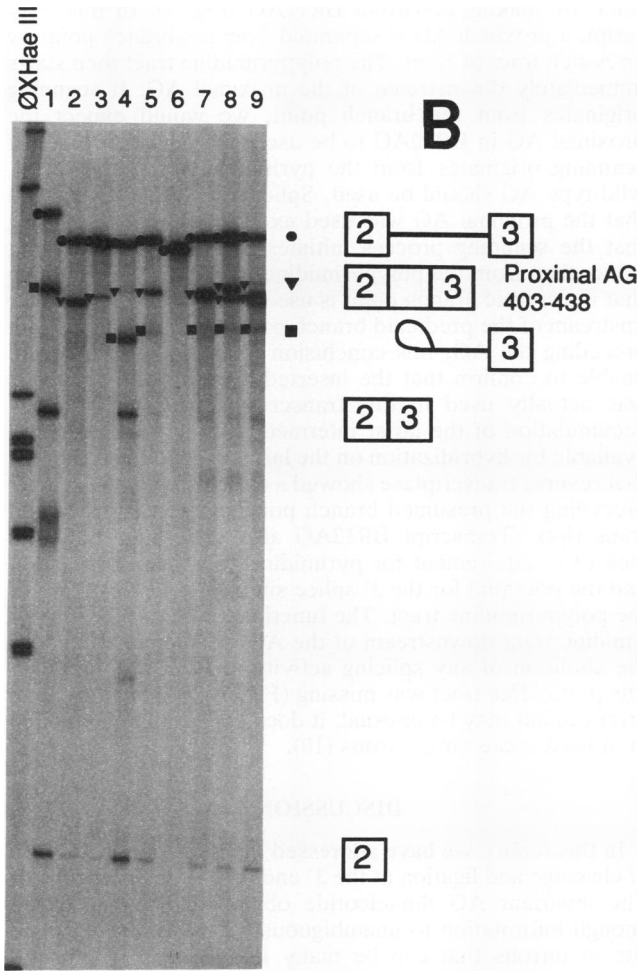




FIG. 7. Effects of branch point-to-AG spacing upon scanning. (A) Schematic representation of transcripts. The parental construct pGC+DX was modified by deleting nt 30 to 110 containing the branch point and polypyrimidine tract sequences (Δ30-110). Oligonucleotides bearing a branch point (underlined) within a strong sequence context, a polypyrimidine tract and one or more AGs (underlined), were reinserted as indicated. To the right of the sequences of the branch point to AG region are shown the sizes of the spliced products if the first or second AG downstream of the branch point is used. Note that for some of the transcripts the second AG is the wild-type CAG, while for others, both the first and second AGs were introduced with the insert. Relative use of the competing AGs (assessed from Fig. 6B) is shown as follows: ++++, complete use of the AG; −, no use of AG; +/−, extremely inefficient use of AG; ++, equal use of first two AGs. (B) In vitro splicing reaction of transcripts shown in panel A. Lanes are numbered as in panel A. Splicing was for 3 h. Bands corresponding to spliced products using the proximal AGs introduced in the oligonucleotides are marked by the triangles.

C over U (72% versus 23% [47]) at the −3 position. Neither the relative competitiveness nor the kinetics of 3′ splice site cleavage, as revealed in the current assay, can explain this preference. These data are consistent with a model in which active scanning stops at the first AG, but when this AG is inefficient and slow in the second step, as is the case for GAG, the spliceosome can still select a more competitive AG within a fixed distance downstream.

The preceding data demonstrate that 3′ splice site specification displays elements of both a scanning-type mechanism and short-range competitive behavior between AGs based upon the preceding nucleotide. Since GAGs in this assay were clearly the least competitive 3′ splice sites, we next investigated whether increasing the distance between a GAG and a downstream CAG could eventually eliminate use of the CAG. To this end we tested a series of transcripts with competing GAGs 12, 30, 50, 75, and 108 nt upstream of the wild-type CAG (Fig. 6B). Because of the slow kinetics of the second step at GAGs, their ability to compete in these transcripts can be monitored only by the reduction in use of the wild-type site. The results clearly show that GAGs at all positions between 12 and 75 nt upstream reduce, but do not completely eliminate, use of the wild-type site. Only the GAG 108 nt upstream appears to cause complete repression of the wild-type CAG. The simplest explanation for this observation is that unlike the other NAGs, which can be used exclusively once they are separated by more than a few nucleotides from the next AG, GAGs are inefficiently recognized, and a proportion of spliceosomes can continue scanning to the next downstream AG. The very weak competitiveness and/or leakiness of GAGs explains their almost

complete exclusion from wild-type 3′ splice sites, since competition from exon AGs would lead to inappropriate splicing.

**Bypassing of AGs very close to the branch point.** Compilations of mapped branch point sequences indicate that very few branch points are closer than ~18 nt to the 3′ splice site (14). This raises the question of whether there is some minimal acceptable separation of branch and 3′ splice site sequences, below which steric hindrance intervenes, perhaps in a manner analogous to that of the minimal functional 5′ splice site-branch point separation (7, 41, 50). This could give rise to the bypassing of proximal AGs. To address this question, we made mutant pΔ30-110 (Fig. 7A), in which the entire branch point-polypyrimidine tract region had been deleted. Synthetic oligonucleotides bearing a strong branch point sequence, a pyrimidine tract, and AG elements in various configurations could then be reinserted into this intron. In this arrangement, the branch point was 69 nt from the 5′ splice site, with no upstream adenosines in a position or sequence context to be used as cryptic branch points.

In mutants B19AG, B12AG, and B4AG a proximal AG was placed 19, 12, or 4 nt downstream of the branch point, with the pyrimidine tract between the two elements, on either side of the AG, or downstream of the AG, respectively (Fig. 7). B19AG was spliced efficiently using exclusively the proximal AG. In mutant B12AG, the proximal AG was still used exclusively, although the efficiency of the reaction was reduced, perhaps because of the discontinuity in the polypyrimidine tract or the proximity of the AG to the branch point. Thus, B19AG and B12AG behave in a manner consistent with simple downstream scanning from the branch point. In contrast, B4AG was spliced efficiently, using almost exclusively the distal wild-type 3′ splice site. These data therefore indicate that AGs as close as 12 nt from the branch can be located by scanning and used exclusively, while a 3′ splice site only 4 nt downstream of the branch is too close to be efficiently recognized, so that scanning then proceeds downstream to the next AG. It is noteworthy, however, that even in B4AG the proximal site was still used to a minor extent, indicating that there is no absolute minimal distance requirement between branch point and 3′ splice site, although recognition may become very inefficient. The use of such closely spaced splicing elements in vertebrate systems is not unprecedented; a 4-nt separation of the branch point and AG also occurs in the polyomavirus middle t intron (10). This suggests that the flexibility of arrangement of 3′ splice site elements encompasses functional pyrimidine tracts downstream of the actual splice site (see also below).

The preceding results showed that an AG only 12 nt from the branch point can be selected and used exclusively when the next downstream AG is ~100 nt distant. In view of the observation that closely spaced AGs are able to compete as 3′ splice sites (see above; Fig. 6) we tested the ability of AGs 19, 12, and 4 nt downstream of the branch point to be used in the presence of a competing AG 24 nt downstream of the branch point (Fig. 7, transcripts B19AG2, B12AG2, and B4AG2). In B19AG2, the proximal AG was still used exclusively, with no use of the AG only 5 nt further downstream, again consistent with simple scanning. In contrast, in B12AG2 the AGs 12 and 24 nt from the branch were used to similar extents. In B4AG2, the AG 24 nt downstream of the branch point was used almost exclusively and use of the more distal wild-type 3′ splice site, used by B4AG, was consequently occluded. Comparison of the results with the B12AG and B12AG2 transcripts indicates that although an AG as close as 12 nt can be recognized efficiently and used exclusively as the 3′ splice site when the next AG is distant (112 nt further downstream in B12AG), a downstream AG can compete efficiently if it is sufficiently close (12 nt further in B12AG2). This is again consistent with the notion that active scanning halts at the first AG but that the spliceosome can still see a limited stretch of downstream RNA, within which a more competitive AG can be used. In the case of B12AG2, the AG 12 nt from the branch point is presumably rendered less competitive because of steric effects arising from its proximity to the branch point, and so the AG 24 nt from the branch can compete. As shown above (Fig. 6), other determinants of competitiveness include the identity of the nucleotide preceding the AG and the distance separating competing AGs.

**Scanning originates from the branch point rather than the pyrimidine tract.** From our previous data we concluded that scanning originates from the branch point-polypyrimidine tract, these two elements being obligatorily linked for the first step in splicing (53). We next decided to test whether scanning originates from the branch point or the pyrimidine tract, by making construct BR12AG (Fig. 7). In this transcript, a proximal AG is separated from the branch point by an A-rich tract of 12 nt. The polypyrimidine tract then starts immediately downstream of the proximal AG. If scanning originates from the branch point, we would expect the proximal AG in BR12AG to be used. On the other hand, if scanning originates from the pyrimidine tract, the distal, wild-type AG should be used. Splicing of BR12AG showed that the proximal AG was used exclusively, demonstrating that the scanning process initiates from the branch point rather than from the polypyrimidine tract. It is worth noting that if a cryptic branch point is used in this transcript, either upstream of the predicted branch point or within the A tract preceding the AG, this conclusion is unaffected. We were unable to confirm that the inserted branch point sequence was actually used in this transcript because of the low accumulation of the lariat intermediate, the very short tail available for hybridization on the lariat product, and the fact that reverse transcriptase showed a strong stop at the A tract preceding the presumed branch point even in unprocessed transcripts. Transcript BR12AG also underlines both the lack of a requirement for pyrimidines upstream of the AG and the potential for the 3′ splice site AG to lie upstream of the polypyrimidine tract. The functional role of this polypyrimidine tract downstream of the AG was demonstrated by the abolition of any splicing activity in BR12ΔY, in which this pyrimidine tract was missing (Fig. 7). Although such an arrangement may be unusual, it does appear to be exploited in at least some viral introns (10).

## DISCUSSION

In this report, we have addressed the issue of how the site of cleavage and ligation at the 3′ ends of introns is specified. The invariant AG dinucleotide obviously does not have enough information to unambiguously identify the 3′ splice site in introns that can be many kilobases in length and contain numerous AG motifs. The associated polypyrimidine tract and branch point sequence elements play a fundamental role in specification of the 3′ splice site. We have previously proposed that the 3′ splice site is located by a simple scanning process from the branch point-polypyrimidine tract, in which the first downstream AG motif is selected. The current experiments lend further support to a scanning model for 3′ splice site recognition. At the same

time, we have found evidence for competition between AGs in the second step in splicing.

**Evidence for a scanning process in 3′ splice site specification.** A number of lines of evidence suggest a form of scanning behavior in the second step in splicing, as opposed to some other basis for preference of proximal AGs. First, we found that CAGs separated by only 6 or 7 nt but located more than 170 nt from the branch point produced exclusive use of only the first AG (Fig. 3A). Second, a stable hairpin structure between the branch point-pyrimidine tract and AG but distant from both elements and therefore unlikely to interfere directly with interactions at those sites blocked the second step in splicing (Fig. 3A) (53). Third, a CAG within the single-stranded loop of a stem-loop structure was not used in the second step in splicing. These results are all consistent with a scanning process but not with a preference for proximal AGs based upon the greater probability of diffusion-controlled collision. The latter process would predict nearly equal use of closely spaced AGs, no effect of hairpins upon use of downstream AGs, and use of proximal AGs in loop regions. In addition, a restricted-diffusion model would not be able to impose the required polarity; if the first downstream AG were distant, simple looping out of intervening RNA might lead to selection of a closer AG upstream of the branch point. Finally, we showed that the exclusive use of proximal AGs does not require that they be preceded by pyrimidine-rich regions (Fig. 2), although we cannot rule out that adjacent pyrimidine-rich regions may have an effect upon competition between AGs similar to that of the immediate preceding nucleotide, as has been demonstrated in yeast cells (32) (see below).

The transcripts that have provided the strongest evidence for scanning-like behavior all have the branch point and pyrimidine tract distant from the 3′ splice site. In such introns the AG has no observable effect on the splicing reaction prior to the second step, while AGs within the more usual 18- to 40-nt distance from the branch point can improve the efficiency of the first step (36). It is pertinent to question whether the findings from such unconventionally arranged introns can be generalized. A number of points can be raised to justify this approach. First, although more-conventional AG-dependent introns show a clear preference for use of the AG proximal to the branch point, it is only by using unconventionally arranged AG-independent introns that this preference can be seen to be due to a scanning-like process rather than some other basis for preferential use of proximal AGs. Second, the arrangement of more conventionally spaced branch point-3′ splice site units and their behavior when mutations introduce new AGs between the branch point and wild-type 3′ splice site (24, 27, 54) are all consistent with the scanning model that we have derived from the α-TM intron. Third, the α-TM branch point-3′ splice site region used in all the transcripts in this study is spliced efficiently in almost all cell types, with the single exception of smooth muscle cells, where its use is specifically repressed (26, 51). Correct 3′ splice site selection within this unusually arranged branch point-3′ splice site region therefore does not require any cell-specific factors but is achieved by the constitutive splicing machinery. It is unlikely that only a minor subset of the splicing machinery in HeLa extracts is competent to engage in a long-range scanning process, since the transcript is spliced efficiently in vitro. If scanning is a general feature of the second step of splicing, it is pertinent to question why the vast majority of 3′ splice sites are arranged such that the scanning behavior is not readily apparent, with the branch point and AG separated by only 18 to 40 nt. We suggest that selective pressure will tend to produce compact splice site arrangements, which are consequently less susceptible to disruption of the spliced message by mutational insertion of AGs between the branch point and true 3′ splice site AG. Only in a few specific cases will an overriding functional constraint tend to maintain a large branch point-to-AG separation. For instance, regulatory constraints underlie the structure of the α-TM intron. The very long pyrimidine tract specifies default selection of exon 3 (26), while other regulatory elements between the polypyrimidine tract and AG are necessary for smooth muscle cell-specific repression of the usually strong default exon (51); these two elements necessarily enforce a large separation of branch point and AG.

In summary, we believe that it is the unusual anatomy of the α-TM intron, with the 3′ splice site AG distant from the branch point, that allows for dissociation between the two steps in splicing and clarifies the scanning-like behavior involved in the second step of the splicing reaction. Equally, transcripts that illustrate the possibility for competition between AGs on the basis of the preceding nucleotide (Fig. 6), steric effects (Fig. 7), or adjacent pyrimidine-purine content (32) utilize uncommon intron arrangements to emphasize the competitive properties of the second step. It is important to note that at this stage, independent physical evidence for spliceosome scanning between the branch point and 3′ splice site is lacking. Nonetheless, the behavior of the transcripts detailed above argues for a scanning process as the simplest explanation for the data.

**Bypassing of and competition between AGs.** We characterized a number of circumstances in which the first AG is not used exclusively. These can be grouped into two main classes. In the first class, the first AG appears not to be recognized at all and the next AG is located by the spliceosome independent of its distance downstream. We identified two such circumstances. The first AG can be bypassed in this way if it is sequestered within a stable secondary structure (Fig. 5) (2, 5) or if it is extremely close to the branch point (Fig. 7, B4AG). The next downstream AG is then used, assuming that the secondary structure is not in a location where it blocks the second step completely (Fig. 4 and 5). In these cases the first AG simply appears to be invisible to the spliceosome.

The second class of bypassed upstream AGs involves competition between AGs. The first AG downstream is recognized, but a more competitive AG downstream can still be used in the second step. The distinguishing feature from the first group of cases is that the second AG must be within a restricted distance downstream of the first in order to be selected: the best examples of such competition involved AGs separated by 3, 6, or 12 nt. The behavior of GAG trinucleotides seems to be intermediate between these two classes, in that they are recognized inefficiently and can reduce but not completely eliminate use of distant downstream AGs.

We identified a number of determinants of competitiveness between AGs.

(i) **Proximity.** The more proximal AG is usually much more competitive, although other determinants of competitiveness may override this effect.

(ii) **Preceding nucleotide.** The identity of the nucleotide preceding the AG affects competitiveness with the hierarchy CAG ≈ UAG > AAG > GAG. An identical hierarchy is observed for the rate of the second step at the different NAGs. This parallels the occupancy of the −3 position in consensus compilations of 3′ splice sites (25, 29, 47) except

that our experimental data show U and C to be equally competitive, whereas C is used four times as frequently as U in 3′ splice sites. In addition to being weakly competitive with nearby AGs, GAGs also appeared to be quite leaky even when the next downstream AG is distant. It seems likely that the identity of the nucleotide following the AG may also play a role in competition; consensus compilations show a clear G preference at the position following the AG (25, 29, 47). The ability of closely spaced AGs to compete on the basis of the nature of the preceding nucleotide suggests that selective pressure may operate upon exon sequences to prevent competition from AGs at the 5′ end of exons. Indeed, in a survey of >6,000 vertebrate exons we have found that there is an area of prohibited competition at the 5′ ends of exons (19). By scoring AGs on the basis of the preceding and following nucleotides it was found that the incidence of potentially highly competitive AGs was prohibited at the 5′ ends of the exons, where inappropriate competition could disrupt the open reading frame. Thus, our experimental data on competition between closely spaced AGs are buttressed by the apparent selective pressure against competitive AGs in exons. A clue to the molecular interactions underlying the consensus C > U > A > G at the position preceding the AG is provided by the recent report that nucleotides C7 and U8 of *Schizosaccharomyces pombe* U1 snRNA base pair with the 3′ splice site AG (38). The equivalent nucleotides C9 and U10 of mammalian U1 snRNAs are followed by G11, which would be expected to have a base pairing potential of C > U > A/G with the nucleotide preceding the AG. We are currently testing the ability of mutants in the U1 position 11 to alter competition between competing NAGs in a reporter construct.

**(iii) Steric effects.** If the first AG is relatively close to the branch (e.g., 12 nt in B12AG2 [Fig. 7]), a slightly more distal AG may be preferred. The poor competitiveness of AGs too close to the branch point probably explains why almost all introns have a branch point-to-AG separation of at least 18 nt (14). The common minimal 18-nt separation most probably therefore results from selective pressure to prevent inappropriate competition from AGs within the 5′ ends of exons rather than from absolute steric interference below this distance.

One potential determinant of competitiveness which we did not thoroughly investigate but which has been shown by Patterson and Guthrie (32) to affect competition between AGs in *Saccharomyces cerevisiae* is the influence of the pyrimidine or purine content upstream of the AG. In that study, two AGs separated by 34 nt, both downstream of a common branch point, competed as 3′ splice sites. An adjacent uridine stretch improved competitive efficiency, while adenosines had the opposite effect. This appears to represent another form of competition between AGs analogous to that determined by the nature of the −3 nucleotide. Although we showed unequivocally that a proximal AG could be used exclusively even when preceded by a purine-rich stretch (Fig. 2), in these transcripts the next CAG was more than 100 nt downstream, beyond the distance where we would expect it to be competitive. A complicating factor that has prevented us from rigorously investigating this potential effect of pyrimidines is the flexibility of 3′ splice site-branch point-pyrimidine tract sequence requirements for the first splicing step in mammalian introns (36). Whereas in yeast cells the branch point UACUAAC sequence is highly conserved and early steps in splicing are absolutely independent of a 3′ splice site AG (43, 44), in mammalian splicing AG independence is conditional upon the strength of the branch point-pyrimidine tract elements (36, 49, 53). Sequences that are not recruited as branch points in the presence of a pyrimidine tract alone may be activated by the same pyrimidine tract followed closely by an AG. Thus, when we placed a proximal AG, preceded by purines, in competition with a downstream AG preceded by a strong pyrimidine tract, use of the downstream AG occurred, due not to a second step competition but to activation of a cryptic branch point (data not shown and legend to Fig. 6). This suggests that competition between mammalian 3′ splice sites based upon the adjacent pyrimidine-purine content may more commonly occur before the first step, involving competing sites of branch formation. Indeed, we have demonstrated that the default selection of α-TM exon 3 results primarily from a competition between exons 2 and 3 based on the relative strengths of the polypyrimidine tracts associated with the two exons (26). Nonetheless, a second-step competition between AGs in some mammalian introns, based upon adjacent pyrimidine-purine content, remains a possibility.

**Parallels with translation initiation?** Our scanning model for 3′ splice site selection bears a number of similarities to the model for AUG scanning in translational initiation, which although not universally employed (18) has been able to accommodate a number of modifications to its simplest formulation (21). Both processes show elements of scanning-like as well as competitive behavior. In both instances the major determinant for selection of a short and therefore relatively unspecific sequence signal is its location as the first such motif downstream of a fixed reference point of much higher specificity. In the case of translation, the 7mG cap is unique, being located only at the 5′ end of the transcript, while in splicing the 3′- to 5′-branched nucleotide within the intron is specified by the invariant UACUAAC sequence in yeast cells or by a looser immediate sequence context and an adjacent polypyrimidine tract in higher eukaryotes. The conclusion that the AG is located with direct reference to the branch point itself, rather than the pyrimidine tract, is demonstrated by the fact that a first AG can be used exclusively as the 3′ splice site even when it lies upstream of the polypyrimidine tract (Fig. 7, BR12AG). Thus, second-step scanning occurs directly between the sites of the two transesterification reactions. In both translational and 3′ splice site scanning, the 5′-most AUG or AG motif, respectively, can sometimes be bypassed for a more distal one if it is very close to the reference signal or if it lies within a poor sequence context. Relative competitiveness is predictable on the basis of a consensus sequence context, GCCGCCRC CAUGG for translation (21) and CAG|G for splicing (25). Proximal AUGs or AGs in weak contexts can be bypassed for more-favorable downstream sites. In addition, both processes can be blocked by the presence of a stable secondary structure between the two signals involved in scanning (Fig. 4 and 5) (20, 34, 41). Indeed, it is striking that *PRP16*, a yeast gene essential for the second step in splicing, has been shown to have regions of homology with the helicase motifs of the translation initiation factor eIF4A (45), which is thought to be important in melting secondary structure in the 5′ end of mRNAs and thus allowing the 40S ribosomal subunit to scan to the first AUG. A number of factors have been implicated in the process of 3′ splice site recognition. These include U2AF (42, 62), IBP (12, 57), PTB (9, 33), and heterogeneous nuclear ribonucleoproteins A1, C, and D (56). However, with the exception of heterogeneous nuclear ribonucleoprotein A1, the polypyrimidine tract rather than the AG appears to be the primary binding

site for all these proteins. In yeast cells, at least four genes, *PRP16* (45), *PRP18* (59), and *SLU4* and *SLU7* (6), have been shown to encode essential second-step splicing factors, which may be candidates for AG recognition factors. However, the recent evidence from yeast that U5 snRNA is able to interact with exon nucleotides flanking both splice sites (28) and that U1 can base pair with the 3′ splice site AG prior to the first step in splicing (38) suggests that RNA-RNA interactions will be critical in locating the AG, as they are in translation initiation via codon-anticodon base pairing (3).

In summary, the data presented here are consistent with a scanning mechanism for 3′ splice site recognition, although competition between AGs also occurs. Exceptions to the rule that the first AG downstream of the branch point is recruited as the 3′ splice site themselves follow predictable rules, as has been found in translational initiation. In this respect mechanisms of translation have provided an illuminating paradigm for yet another aspect of the splicing reaction.

## ACKNOWLEDGMENTS

We thank Emily Flynn McIntosh for preparing the figures, Ok Hwang for sequencing a number of the constructs, Jim Patton and Roger Breitbart for critical comments upon the manuscript, and other members of our departments for helpful suggestions and insights. We also thank Steen Knudsen for providing information on computer analysis of gene sequences.

This work was supported in part by grants from the National Institutes of Health and the Muscular Dystrophy Association of America (to B.N.-G.) and from the Nuffield Foundation (to C.W.J.S.).

## REFERENCES

1. **Abmayr, S. M., R. Reed, and T. Maniatis.** 1988. Identification of a functional mammalian spliceosome containing unspliced pre-mRNA. Proc. Natl. Acad. Sci. USA **85:**7216–7220.
2. **Chebli, K., R. Gattoni, P. Schmitt, G. Hildwein, and J. Stevenin.** 1989. The 216-nucleotide intron of the E1a pre-mRNA contains a hairpin structure that permits utilization of unusually distant branch acceptors. Mol. Cell. Biol. **9:**4852–4861.
3. **Cigan, A. M., L. Feng, and T. F. Donahue.** 1988. $tRNAi^{met}$ functions in directing the scanning ribosome to the start site of translation. Science **242:**93–97.
4. **Cook, K. S., D. L. Groves, H. Y. Min, and B. M. Spiegelman.** 1985. A developmentally regulated mRNA from 3T3 adipocytes encodes a novel serine protease homologue. Proc. Natl. Acad. Sci. USA **82:**6480–6484.
5. **Deshler, J. O., and J. J. Rossi.** 1991. Unexpected point mutations activate cryptic 3′ splice sites by perturbing a natural secondary structure within a yeast intron. Genes Dev. **5:**1252–1263.
6. **Frank, D., B. Patterson, and C. Guthrie.** 1992. Synthetic lethal mutations suggest interactions between U5 small nuclear RNA and four proteins required for the second step of splicing. Mol. Cell. Biol. **12:**5197–5205.
7. **Fu, X.-Y., J. D. Colgan, and J. L. Manley.** 1988. Multiple *cis*-acting sequence elements are required for efficient splicing of simian virus 40 small-t antigen pre-mRNA. Mol. Cell. Biol. **8:**3582–3590.
8. **Fu, X.-Y., H. Ge, and J. L. Manley.** 1988. The role of the polypyrimidine stretch at the SV40 early pre-mRNA 3′ splice site in alternative splicing. EMBO J. **7:**809–817.
9. **Garcia-Blanco, M. A., S. F. Jamison, and P. A. Sharp.** 1989. Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. Genes Dev. **3:**1874–1886.
10. **Ge, H., J. Noble, J. Colgan, and J. L. Manley.** 1990. Polyoma virus small tumor antigen pre-mRNA splicing requires cooperation between two 3′ splice sites. Proc. Natl. Acad. Sci. USA **87:**3338–3342.
11. **Ge, H., P. Zuo, and J. L. Manley.** 1991. Primary structure of the human splicing factor ASF reveals similarities with Drosophila regulators. Cell **66:**373–382.
12. **Gerke, V., and J. A. Steitz.** 1986. A protein associated with small nuclear ribonucleoprotein particles recognizes the 3′ splice site of premessenger RNA. Cell **47:**973–984.
13. **Goux-Pelletan, M., D. Libri, Y. d'Aubenton-Carafa, M. Fiszman, E. Brody, and J. Marie.** 1990. In vitro splicing of mutually exclusive exons from the chicken β-tropomyosin gene: role of the branch point location and very long pyrimidine stretch. EMBO J. **9:**241–249.
14. **Green, M. R.** 1986. Pre-mRNA splicing. Annu. Rev. Genet. **20:**671–708.
15. **Halfter, H., and D. Gallwitz.** 1988. Impairment of yeast pre-mRNA splicing by potential secondary structure forming sequences near the conserved branchpoint sequence. Nucleic Acids Res. **16:**10413–10423.
16. **Helfman, D. M., and W. M. Ricci.** 1989. Branch point selection in alternative splicing of tropomyosin pre-mRNAs. Nucleic Acids Res. **17:**5633–5650.
17. **Helfman, D. M., R. F. Roscigno, G. J. Mulligan, L. A. Finn, and K. S. Wever.** 1990. Identification of two distinct intron elements involved in alternative splicing of β-tropomyosin pre-mRNA. Genes Dev. **4:**98–110.
18. **Jackson, R. J., M. T. Howell, and A. Kaminski.** 1990. The novel mechanism of initiation of picornavirus RNA translation. Trends Biochem. Sci. **12:**477–483.
19. **Knudsen, S., and C. W. J. Smith.** Unpublished observations.
20. **Kozak, M.** 1986. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc. Natl. Acad. Sci. USA **83:**2850–2854.
21. **Kozak, M.** 1989. The scanning model for translation: an update. J. Cell Biol. **108:**229–241.
22. **Langford, C. J., and D. Gallwitz.** 1983. Evidence for an intron contained sequence required for the splicing of yeast RNA polymerase II transcripts. Cell **33:**519–527.
23. **Maniatis, T., and R. Reed.** 1987. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. Nature (London) **325:**673–678.
24. **Metherall, J. E., F. S. Collins, J. Pan, S. M. Weissman, and B. G. Forget.** 1986. β° thalassemia caused by a base substitution that creates an alternative splice acceptor site in an intron. EMBO J. **5:**2551–2557.
25. **Mount, S. M.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10:**459–472.
26. **Mullen, M. P., C. W. J. Smith, J. G. Patton, and B. Nadal-Ginard.** 1991. α-Tropomyosin mutually exclusive exon selection: competition between branch point/polypyrimidine tracts determines exon choice. Genes Dev. **5:**642–655.
27. **Nakano, T., and K. Suzuki.** 1989. Genetic cause of a juvenile form of Sandhoff disease. J. Biol. Chem. **264:**5155–5158.
28. **Newman, A. J., and C. Norman.** 1992. U5snRNA interacts with exon sequences at 5′ and 3′ splice sites. Cell **68:**743–754.
29. **Ohshima, Y., and Y. Gotoh.** 1987. Signals for the selection of a splice site in pre-mRNA. J. Mol. Biol. **195:**247–259.
30. **Padgett, R. A., P. J. Grabowski, M. M. Konarska, S. Seiler, and P. A. Sharp.** 1986. Splicing of messenger RNA precursors. Annu. Rev. Biochem. **55:**1119–1150.
31. **Parker, R., P. G. Siliciano, and C. Guthrie.** 1987. Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. Cell **49:**229–239.
32. **Patterson, B., and C. Guthrie.** 1991. A U-rich tract enhances usage of an alternative 3′ splice site in yeast. Cell **64:**181–187.
33. **Patton, J. G., S. A. Mayer, P. Tempst, and B. Nadal-Ginard.** 1991. Characterization and molecular cloning of polypyrimidine tract binding protein: a component of a complex necessary for pre-mRNA splicing. Genes Dev. **5:**1237–1251.
34. **Pelletier, J., and N. Sonenberg.** 1985. Insertion mutagenesis to increase secondary structure within the 5′ noncoding region of a eukaryotic mRNA reduces translational efficiency. Cell **40:**515–526.
35. **Pikielny, C. W., and M. Rosbash.** 1985. mRNA splicing efficiency in yeast and the contribution of nonconserved se-

quences. Cell **41**:119–126.
36. **Reed, R.** 1989. The organization of 3′ splice site sequences in mammalian introns. Genes Dev. **3**:2113–2123.
37. **Reed, R., and T. Maniatis.** 1985. Intron sequences involved in lariat formation during pre-mRNA splicing. Cell **41**:95–105.
38. **Reich, C. I., R. W. VanHoy, G. L. Porter, and J. A. Wise.** 1992. Mutations at the 3′ splice site can be suppressed by compensatory base changes in U1 snRNA in fission yeast. Cell **69**:1159–1169.
39. **Ruby, S. W., and J. Abelson.** 1991. Pre-mRNA splicing in yeast. Trends Genet. **7**:79–85.
40. **Ruskin, B., and M. R. Green.** 1985. Role of the 3′ splice site consensus sequence in mammalian pre-mRNA splicing. Nature (London) **317**:732–734.
41. **Ruskin, B., J. M. Greene, and M. R. Green.** 1984. Cryptic branch point activation allows accurate in vitro splicing of human β-globin intron mutants. Cell **41**:833–844.
42. **Ruskin, B., P. D. Zamore, and M. R. Green.** 1988. A factor, U2AF, is required for U2snRNP binding and splicing complex assembly. Cell **52**:207–219.
43. **Rymond, B. C., and M. Rosbash.** 1985. Cleavage of 5′ splice site and lariat formation are independent of 3′ splice site in yeast mRNA splicing. Nature (London) **317**:735–737.
44. **Rymond, B. C., D. D. Torrey, and M. Rosbash.** 1987. A novel role for the 3′ region of introns in pre-mRNA splicing of Saccharomyces cerevisiae. Genes Dev. **1**:238–246.
45. **Schwer, B., and C. Guthrie.** 1991. PRP16 is an RNA-dependent ATPase that interacts transiently with the spliceosome. Nature (London) **349**:494–499.
46. **Seraphin, B., L. Kretzner, and M. Rosbash.** 1988. A U1 snRNA: pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5′ cleavage site. EMBO J. **7**:2533–2538.
47. **Shapiro, M. B., and P. Senepathy.** 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. **15**:7155–7174.
48. **Siliciano, P. G., and C. Guthrie.** 1988. 5′ splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. Genes Dev. **2**:1258–1267.
49. **Smith, C. W. J.** Unpublished observations.
50. **Smith, C. W. J., and B. Nadal-Ginard.** 1989. Mutually exclusive splicing of α-tropomyosin exons enforced by an unusual lariat branch point location; implications for constitutive splicing. Cell **56**:749–758.
51. **Smith, C. W. J., and B. Nadal-Ginard.** Unpublished observations.
52. **Smith, C. W. J., J. G. Patton, and B. Nadal-Ginard.** 1989. Alternative splicing in the control of gene expression. Annu. Rev. Genet. **23**:527–577.
53. **Smith, C. W. J., E. B. Porro, J. G. Patton, and B. Nadal-Ginard.** 1989. Scanning from an independently specified branch point defines the 3′ splice site of mammalian introns. Nature (London) **342**:243–247.
54. **Spritz, R. A., P. Jagardeeswaran, P. V. Choudary, P. A. Biro, J. T. Elder, J. K. deRiel, J. L. Manley, M. L. Gefter, B. G. Forget, and S. M. Weissman.** 1981. Base substitutions in an intervening sequence of a $\beta^+$-thalassemia human globin gene. Proc. Natl. Acad. Sci. USA **78**:2455–2459.
55. **Sun, X.-H., and D. Baltimore.** 1991. An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell **64**:459–470.
56. **Swanson, M. S., and G. Dreyfuss.** 1988. RNA binding specificity of hnRNP proteins: a subset bind to the 3′ end of introns. EMBO J. **7**:3519–3529.
57. **Tazi, J., C. Alibert, J. Temsamani, I. Reveillaud, G. Cathala, C. Brunel, and P. Jeanteur.** 1986. A protein that specifically recognizes the 3′ splice site of mammalian pre-mRNA introns is associated with a small nuclear ribonucleoprotein. Cell **47**:755–766.
58. **Ulfendahl, P. J., J.-P. Kreivi, and G. Akusjarvi.** 1989. Role of the branch site/3′-splice site region in adenovirus-2 E1A pre-mRNA alternative splicing: evidence for 5′- and 3′-splice site cooperation. Nucleic Acids Res. **17**:925–938.
59. **Vijayraghavan, U., and J. Abelson.** 1990. PRP18, a protein required for the second reaction in pre-mRNA splicing. Mol. Cell. Biol. **10**:324–332.
60. **Wassarman, D. A., and J. A. Steitz.** 1992. Interactions of small nuclear RNA's with precursor messenger RNA during in vitro splicing. Science **257**:1918–1925.
61. **Wu, J., and J. L. Manley.** 1989. Mammalian pre-mRNA branch site selection by U2snRNP involves base pairing. Genes Dev. **3**:1553–1561.
62. **Zamore, P. D., and M. R. Green.** 1989. Identification, purification and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc. Natl. Acad. Sci. USA **86**:9243–9247.
63. **Zhuang, Y., and A. M. Weiner.** 1986. A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. Cell **46**:827–835.
64. **Zhuang, Y., and A. M. Weiner.** 1989. A compensatory base change in human U2 snRNA can suppress a branch site mutation. Genes Dev. **3**:1545–1552.