MOLECULAR AND CELLULAR BIOLOGY, June 1990, p. 2492–2502 Vol. 10, No. 6
0270-7306/90/062492-11$02.00/0

Structure and Expression of the *Drosophila melanogaster* Gene for the U1 Small Nuclear Ribonucleoprotein Particle 70K Protein

RICARDO MANCEBO, PATRICK C. H. LO, AND STEPHEN M. MOUNT*

Department of Biological Sciences, Columbia University, New York, New York 10027

Received 11 January 1990/Accepted 14 February 1990

A genomic clone encoding the *Drosophila* U1 small nuclear ribonucleoprotein particle 70K protein was isolated by hybridization with a human U1 small nuclear ribonucleoprotein particle 70K protein cDNA. Southern blot and in situ hybridizations showed that this U1 70K gene is unique in the *Drosophila* genome, residing at cytological position 27D1,2. Polyadenylated transcripts of 1.9 and 3.1 kilobases were observed. While the 1.9-kilobase mRNA is always more abundant, the ratio of these two transcripts is developmentally regulated. Analysis of cDNA and genomic sequences indicated that these two RNAs encode an identical protein with a predicted molecular weight of 52,879. Comparison of the U1 70K proteins predicted from *Drosophila*, human, and *Xenopus* cDNAs revealed 68% amino acid identity in the most amino-terminal 214 amino acids, which include a sequence motif common to many proteins which bind RNA. The carboxy-terminal half is less well conserved but is highly charged and contains distinctive arginine-rich regions in all three species. These arginine-rich regions contain stretches of arginine-serine dipeptides like those found in *transformer*, *transformer-2*, and *suppressor-of-white-apricot* proteins, all of which have been identified as regulators of mRNA splicing in *Drosophila melanogaster*.

Splicing of pre-mRNA occurs in a two-step reaction requiring the assembly of a large complex (the spliceosome) which includes several small nuclear ribonucleoprotein particles (snRNPs) (see references 19, 32, and 52 for reviews). Each snRNP consists of at least one snRNA bound to several proteins, some of which are common to all the spliceosomal snRNPs. The human U1 snRNP is composed of a 164-nucleotide (nt) U1 snRNA molecule complexed to at least 10 proteins (30). It binds to 5′ splice site sequences by base pairs involving the 5′ end of U1 RNA (12, 24, 35, 62). There is also evidence that the U1 snRNP interacts with the U2 snRNP during splicing (8, 12), and that this interaction occurs early and is required for further stages of splicing (25, 48, 50). Three of the human U1 snRNP proteins, 70K, A, and C, are unique to the U1 snRNP (21, 43); the other U1 snRNP proteins are common to the U2, U5, and U4/U6 snRNPs.

Patients with autoimmune disorders produce antibodies which specifically immunoprecipitate the U1 snRNP (28), and many of these are directed against epitopes on the 70K protein (43, 60). Anti-RNP antisera have been shown to block in vitro splicing when they are added to splicing reactions (40). Such sera, as well as a mouse monoclonal antibody to the mouse U1 snRNP 70K protein (6), have enabled researchers to clone cDNAs encoding the U1 70K protein from expression libraries (47, 59).

Sequence data from these cDNAs have provided information regarding the primary structure of the U1 70K protein. First, it appears that this protein is actually 52 kilodaltons in size (45, 53). We have followed others (15, 46) in using U1 70K or 70K as a name for this protein, although this name says nothing about the size of the protein. A conserved region of amino acid sequence found in a number of RNA-binding proteins (1, 3, 33, 57) occurs in the amino-terminal half of the human U1 70K protein at amino acids 104 to 183, and this region of the protein together with a minimum of flanking amino acids (amino acids 92 to 202) is capable of binding U1 snRNA specifically (45). The human U1 70K protein makes contacts with U1 snRNA at the first stem-loop (20, 41) in a sequence-specific manner (46, 56). The carboxy-terminal portion of the protein is particularly rich in arginine residues and contains repetitions of arginine-glutamic acid, arginine-aspartic acid, and arginine-serine. It appears that this region is responsible for the altered mobility of the 70K protein on sodium dodecyl sulfate-polyacrylamide gels (45). Similar charged regions have been found in the predicted protein products of the *Drosophila suppressor-of-white-apricot* [$su(w^a)$] (13), *transformer* (*tra*) (9), and *transformer-2* (*tra-2*) (18) loci, as well as in a mouse major histocompatibility complex gene of unknown function (29). All these genes with the exception of the last have recently been shown to be regulators of RNA splicing (2, 61).

The U1 70K protein is ideally suited among spliceosomal components for a detailed functional investigation for several reasons. Among them are the availability of monospecific antisera, the presence of the 70K protein in a complex (the U1 snRNP) having a known and important function in splicing, and the existence of a sequence motif found in proteins which act to regulate splicing. The precise role of the U1 70K protein in splicing is unknown. Specific regions of this protein could participate in (i) the recognition of the 5′ splice site by the U1 snRNP; (ii) interactions between U1 and U2 snRNPs; (iii) interactions between U1 snRNPs and other proteins required for splicing; (iv) interactions between U1 snRNPs and other RNAs. We have chosen *Drosophila melanogaster* as a system in which to analyze U1 70K protein function using genetic and biochemical techniques. In this communication, we report on the structure and expression of the *Drosophila* U1 70K protein gene. Of particular interest is the pattern of conservation of sequence elements within the protein-coding region.

MATERIALS AND METHODS

Screening of genomic and cDNA libraries. To isolate *Drosophila* U1 70K genomic clones, we screened a λEMBL4 genomic library (Oregon R strain; prepared by Mike Goldberg) with the human 70K cDNA clone (provided by C. C.

* Corresponding author.

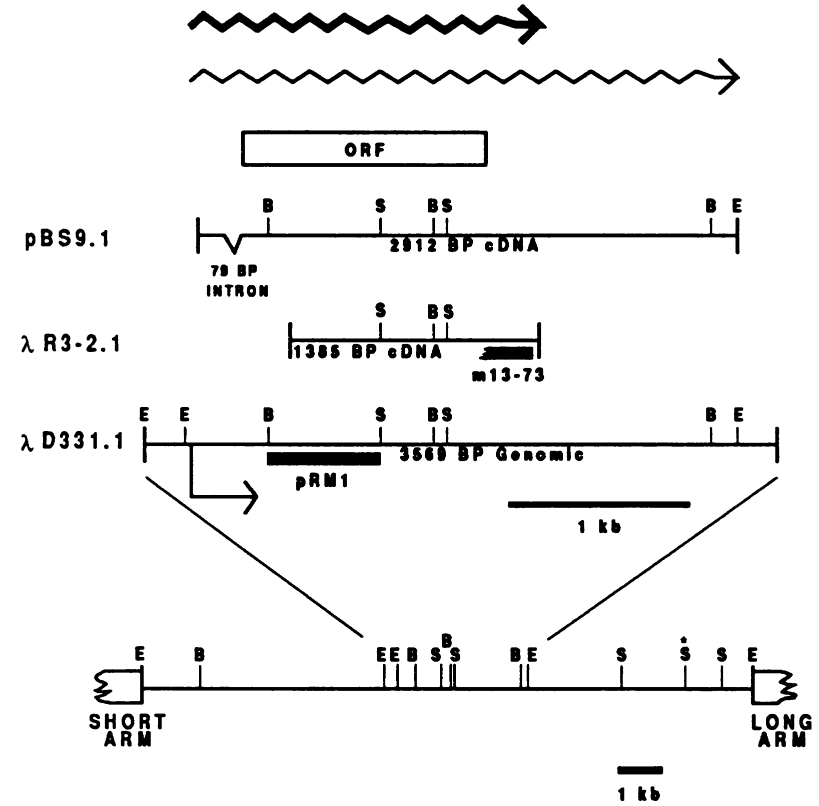

FIG. 1. Structure of the cDNA and genomic clone inserts. The bold and thin wavy lines depict the major and minor U1 70K gene transcripts, respectively, and their direction of transcription. ORF is the 448-amino-acid ORF determined by conceptual translation of the 2.9-kb insert within pBS9.1, a cDNA clone. A 79-bp intron lines 48 bp upstream of the ORF in pBS9.1. λR3-2.1 has a 1.4-kb cDNA insert extending to a natural polyadenylation site. The position of the subclone used to isolate pBS9.1 is shown beneath the map of λR3-2.1. The jagged end of m13-73 indicates that this end has not been mapped. The bottom line shows a restriction map of the entire 14.5-kb genomic insert of λD331.1. Above it is an enlargement of the λD331.1 region from which the cDNAs are derived. The arrow below this enlargement indicates again the start and direction of transcription of the U1 70K gene. pRM1 is a 620-bp *Bam*HI-*Sal*I genomic fragment in pUC19. Enzyme sites are as follows: E, *Eco*RI, B, *Bam*HI; S, *Sal*I. The star denotes one of two possible positions for this *Sal*I site.

Query and J. D. Keene) by the method of Maniatis et al. (31). The probe was labeled as described by Feinberg and Vogelstein (17).

For cDNA cloning, a total of approximately 100,000 plaques from 12- to 24-h-old embryonic, late-third-instar larva, and adult male libraries in λgt10 (44) were screened with a subclone of λD331.1 in M13 (the 620-nt *Bam-Sal* fragment of pRM1; Fig. 1) by the method of Maniatis et al. (31). From an adult head cDNA library prepared by selecting fragments larger than 2.3 kilobases (kb) in the Lambda Zap vector (Stratagene, La Jolla, Calif.), 270,000 plaques were screened with m13-73 (Fig. 1) as a probe. The probe was labeled as described by Feinberg and Vogelstein (17).

Southern and Northern (RNA) blots. Southern blots were done by the method of Maniatis et al. (31) by loading 5 μg of Oregon R genomic DNA digested with *Bam*HI, *Eco*RI, *Hin*dIII, *Pst*I, or *Sal*I onto a 0.8% agarose gel. The radioactive probe (pBS9.1; Fig. 1) was synthesized as described by Feinberg and Vogelstein (17).

For Northern blots, 5 μg of poly(A)$^+$ mRNA from the indicated stages of development was electrophoretically separated on a 1.2% agarose-formaldehyde gel, soaked in 20× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) for 30 min, and transferred to a GeneScreen Plus (Dupont, NEN Research Products, Boston, Mass.) membrane for 16 to 20 h. A single-stranded radioactive probe was synthesized by annealing a synthesized 17-mer to clone m13-73 (Fig. 1) and incorporating [α-^{32}P]dCTP with the Klenow fragment. The probe was heat denatured and separated on a low-melting-temperature agarose gel (see reference 39).

In situ hybridizations. In situ hybridizations were done by the method of Laverty and Rubin (personal communication and unpublished data).

Isolation of poly(A)$^+$ RNA. Poly(A)$^+$ RNA was isolated from embryonic, larval, pupal, and adult stages of Oregon R flies as described previously (39), using a Brinkmann Polytron homogenizer.

DNA sequencing. Random subfragments of the pBS9.1 and λR3-2.1 inserts were generated by sonication, subcloned into M13mp18, and sequenced by the dideoxynucleotide-chain termination method (49) with the modified T7 DNA polymerase Sequenase (U.S. Biochemical Corp., Cleveland, Ohio) (58). Underrepresented cDNA regions and genomic sequence were directly cloned into the Bluescript SK (M13−) vector (Stratagene) and sequenced.

The U1 70K cDNA sequences were analyzed with the Beckman Microgenie sequence analysis software. Protein comparisons were done with the FASTA program (42) to search sequences translated from an updated GenBank database (version 61) with the derived *Drosophila* U1 70K protein sequence.

Nuclease S1 analysis. To make an antisense probe, a standard 17-mer (1211; New England BioLabs, Inc., Beverly, Mass.) was annealed to a clone (m13RM1) which contains a 166-nt *Bam*HI-*Sca*I U1 70K fragment cloned into a *Bam*HI-*Hin*cII-cut M13mp18 vector and the nontranscribed strand was extended with the Klenow fragment in the presence of 70 μM each cold deoxynucleoside triphosphate. The reaction was stopped, and the double-stranded DNA was then cut with *Bam*HI. The *Bam*HI 5′ overhang was end filled in the presence of 5.5 μM each [α-^{32}P]dGTP and [α-^{32}P]dATP and 500 μM each ddTTP and ddCTP with the Klenow fragment. The probe was heat denatured and isolated on a 6% polyacrylamide–8.3 M urea gel. A total of 12,000 cpm were recovered after eluting the probe from the gel and were divided among four separate reactions. Hybridizations to 12 to 30 μg of poly(A)$^+$ RNA were done in 80% formamide at 30°C for 16 to 20 h. The hybridizations were combined into a single tube and digested with S1 nuclease at a concentration of 15 U of enzyme per μg of RNA for 2, 10, and 50 min at 37°C. Reactions were stopped with 4 M ammonium acetate–20 mM EDTA and the mixtures were extracted with phenol-chloroform before loading onto a 6% polyacrylamide–8.3 M urea gel. The dried gel was exposed with Kodak XOmat-AR with a Quanta III screen at −80°C for 5 days.

Enzymes, nucleotides, and primers. Bacterial restriction enzymes, the Klenow fragment, T4 DNA ligase, and T4 polynucleotide kinase were purchased from New England BioLabs. Sequenase and T4 DNA polymerase were purchased from U.S. Biochemical Corp. S1 nuclease and reverse transcriptase were purchased from Boehringer Mannheim Biochemicals (Indianapolis, Ind.). Radioactive nucleotide triphosphates [α-^{35}S]dATP and [α-^{32}P]dCTP were purchased from Dupont, NEN Research Products. [α-^{32}P] dGTP and [α-^{32}P]dATP were purchased from ICN Radiochemicals. [γ-^{32}P]dATP and nucleotide triphosphates were purchased from Amersham Corp. (Arlington Heights, Ill.) and Pharmacia, Inc. (Piscataway, N.J.), respectively. Primers were purchased from Pharmacia and U.S. Biochemical Corp.

RESULTS

Isolation of *Drosophila* U1 70K genomic and cDNA clones. Because human anti-(U1) RNP antisera specifically recognize the *Drosophila* U1 snRNP (36), we expected the U1 70K protein to be highly conserved. Therefore, we attempted to isolate the *Drosophila* U1 70K protein gene by hybridization to a human U1 70K cDNA clone. A genomic library of *Drosophila* DNA in λEMBL4 was screened with a human U1 70K cDNA clone (45) as a probe. These hybridizations were done under standard conditions. However, from approximately 120,000 plaques (10 genome equivalents) screened, only one clone showed strong hybridization to the human probe. The restriction map of this clone, designated λD331.1, is shown in Fig. 1. A 620-base-pair (bp) *Bam*HI-*Sal*I fragment which hybridized strongly to the human probe was subcloned, and preliminary sequencing confirmed that this *Drosophila* clone contained sequences related to the gene for the human U1 70K protein.

Four different cDNA libraries were screened to isolate *Drosophila* U1 70K cDNA clones. In the first screen, a subclone of λD331.1 in M13 (the 620-nt *Bam-Sal* fragment of pRM1; Fig. 1) was used to screen adult male, late-third-instar larva, and 12- to 24-hour-old embryonic cDNA libraries (44). One clone (λR3-2.1) with a 1.4-kb insert was isolated from the adult male library, and a second clone (λI4-2.9) with a 1.1-kb insert was isolated from the late-third-instar larva library. Because a larger transcript had also been seen in Northern blots (see below), we sought to obtain clones corresponding to that RNA. For this purpose, we used an M13 subclone of λR3-2.1 (m13-73; Fig. 1) to screen a library made from poly(A)$^+$ RNA selected to be larger than 2.3 kb. Five cDNA clones with inserts between 2.4 and 2.9 kb were isolated. The largest of these, pBS9.1, was used in most of the subsequent analyses. In these screenings, we consistently obtained about one U1 70K cDNA per 50,000 plaques.

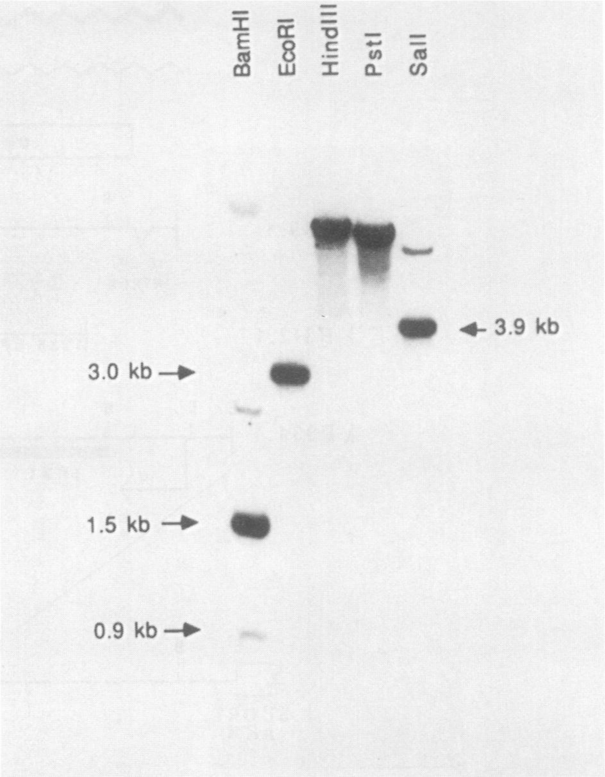

FIG. 2. Southern analysis of the *Drosophila* U1 70K gene. A nitrocellulose filter blot containing 5 μg of Oregon R DNA per lane was hybridized with ^{32}P-labeled pBS9.1. The restriction enzymes used for each digest are indicated. Note that lanes 2, 3, and 4 show only a single hybridizing fragment.

***Drosophila* genome contains only a single U1 70K gene.** To determine the copy number of the *Drosophila* U1 70K protein gene, Oregon R adult DNA was digested with the indicated enzymes (Fig. 2) and probed with our largest cDNA (pBS9.1). A single hybridizing band is seen in lanes 2, 3, and 4. The size of the *Eco*RI fragment of 3.0 kb is as expected from the restriction map of the λD331.1 genomic clone. Digestion with *Sal*I also generated a strongly hybridizing band of the expected size (3.9 kb). For *Bam*HI, the expected number of bands (four) is seen, and those whose size is precisely as in the bacteriophage clone are indicated. However, an unexpected band of approximately 2.6 kb is present and is probably due to polymorphism in a region either upstream or downstream of the 70K gene. Otherwise, all the fragments seen in Fig. 2 are precisely as expected from the structure of the genomic clone. The use of λD331.1 as a probe on an identical blot (data not shown) revealed hybridization to all the bands detected with pBS9.1 and a large number of additional bands, presumably due to the presence of repeated sequences within the insert but outside of the region of the cDNA. These results show that the U1

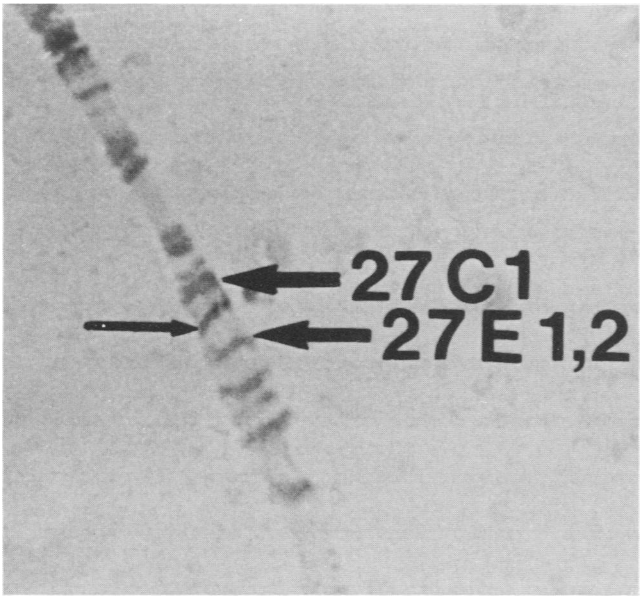

FIG. 3. Cytological localization of the *Drosophila* U1 70K gene. Polytene chromosomes were hybridized to biotin-labeled pRM1. A single band of hybridization was seen at position 27D1,2 on the left arm of chromosome 2 (thin arrow). The position of hybridization is obvious on the original slide owing to a color difference between the chromosomes and the reaction product which does not reproduce well in black and white.

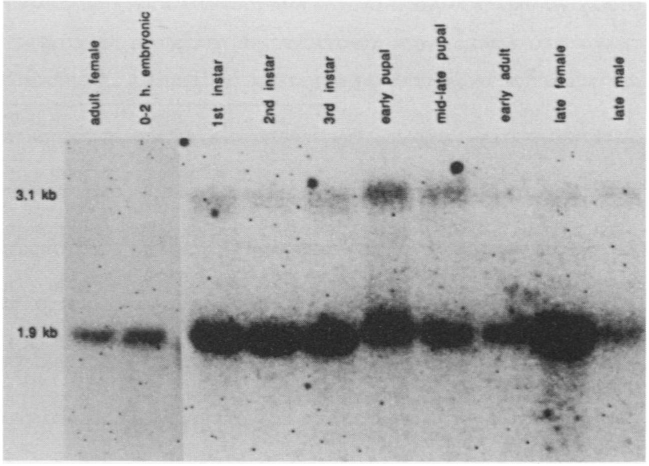

FIG. 4. Expression of poly(A)$^+$ U1 70K RNAs through development. Poly(A)$^+$ RNA was isolated from the designated stages, and 5 μg per lane was electrophoresed on a 1.2% agarose-formaldehyde gel, blotted, and hybridized to a single-stranded cDNA probe (m13-73).

70K protein gene is present in only a single copy in the *Drosophila* genome. This result was confirmed, and the cytological position of the U1 70K genes was established, by hybridization to polytene chromosomes isolated from third-instar Oregon R larvae. Using pRM1 as a probe, a single band of hybridization was observed at 27D1,2, on the left arm of chromosome 2 (Fig. 3).

Two poly(A)$^+$ RNAs of 3.1 and 1.9 kb vary in abundance through development. To determine the structure and temporal pattern of expression of the U1 70K gene, poly(A)$^+$ RNA from different stages of the *Drosophila* life cycle was fractionated on formaldehyde-agarose gels, transferred to filters, and hybridized to the single-stranded probe m13-73 (Fig. 1). Two bands of approximately 3.1 and 1.9 kb were observed (Fig. 4). The smaller RNA was much more abundant at all stages. However, the ratio of the 3.1-kb mRNA to the 1.9-kb mRNA varied through development. The greatest amount of the large species was observed in RNA from pupal stages, and the least was observed in 0- to 2-h-old embryos and adult females. Densitometer readings (see Materials and Methods) indicated that the ratio of the 3.1-kb to the 1.9-kb mRNA in larval stages (Fig. 4, lanes 3, 4 and 5) is between 1:10 and 1:5. In pupae (lanes 6 and 7), the ratio increased about twofold. However, in newly eclosed adults, the ratio was similar to that found in larvae (lane 8). Considerably less of the larger RNA was seen in adult females of egg-laying age (lanes 1 and 9), while identically aged adult males (lane 10) had a ratio like that seen in larvae and recently eclosed adults. This difference is most likely due to maternal deposition of predominantly (or exclusively) the smaller RNA in oocytes. Our observation (lane 2) that the 3.1-kb RNA was nearly undetectable in early embryos is consistent with this.

Gene structure. The 1.4- and 2.9-kb cDNA clone inserts were sequenced by the method of Sanger et al. (49). The sequence of the 1.4-kb insert in λR3-2.1 was identical to the corresponding region in the larger pBS9.1 clone, except for the presence of a short polyadenine tract at the 3′ end, beginning at nt 1932 (see dot in Fig. 5) and 11 nt downstream of an AATAAA stretch. Thus, the λR3-2.1 insert extends from position 547 to a polyadenylation site at position 1932. In contrast, the pBS9.1 insert extends from position 37 to include sequences well beyond position 1932. All the five cDNA clones isolated from the size-selected library were truncated at a natural *Eco*RI site well downstream of the first polyadenylation site (position 3022, Fig. 5). Because the size of the larger RNA seen on Northern blots closely matches the size of this insert in pBS9.1, and the *Eco*RI site occurs 5 nt downstream of the polyadenylation signal sequence AATAAA, we suspected that the polyadenylation site of the larger RNA would lie immediately downstream of the *Eco*RI site.

To map the 3′ end of the larger RNA precisely, we did S1 nuclease protection analysis using a 3′-end-labeled single-stranded fragment derived from an M13 subclone. This fragment contains 41 nt of M13 polylinker at the 5′ end and 166 nt corresponding to a *Bam*HI-*Sca*I fragment at positions 2880 to 3045 (Fig. 5). When this probe was hybridized to poly(A)$^+$ RNA from second-instar larvae and digested with S1 nuclease for different times, the clearest S1 protection product (seen in lane 10′ of Fig. 6) was 160 nt, indicating a polyadenylation site at position 3040, 13 bases downstream of the *Eco*RI site and 24 bases downstream of AATAAA. Because this polyadenylation site falls close to the end of sequence identity between the probe and the gene, a second S1 protection analysis was done with a second probe that extended from the *Bam*HI site at position 2880 to position 3242 (Fig. 5) and contained an additional 57 bases of polylinker sequence at the 5′ end. This second analysis confirmed the results of the first analysis (unpublished data).

The numbering of nucleotides in Fig. 5 is based on a putative start site which lies 30 nt downstream of the sequence TATATTTA and matches the *Drosophila* cap site consensus (22). This start site was indicated by a primer extension product generated with a primer in exon 1 (unpublished data). A second primer gave inconclusive results. We are investigating the structure of the 5′ end of the mRNA

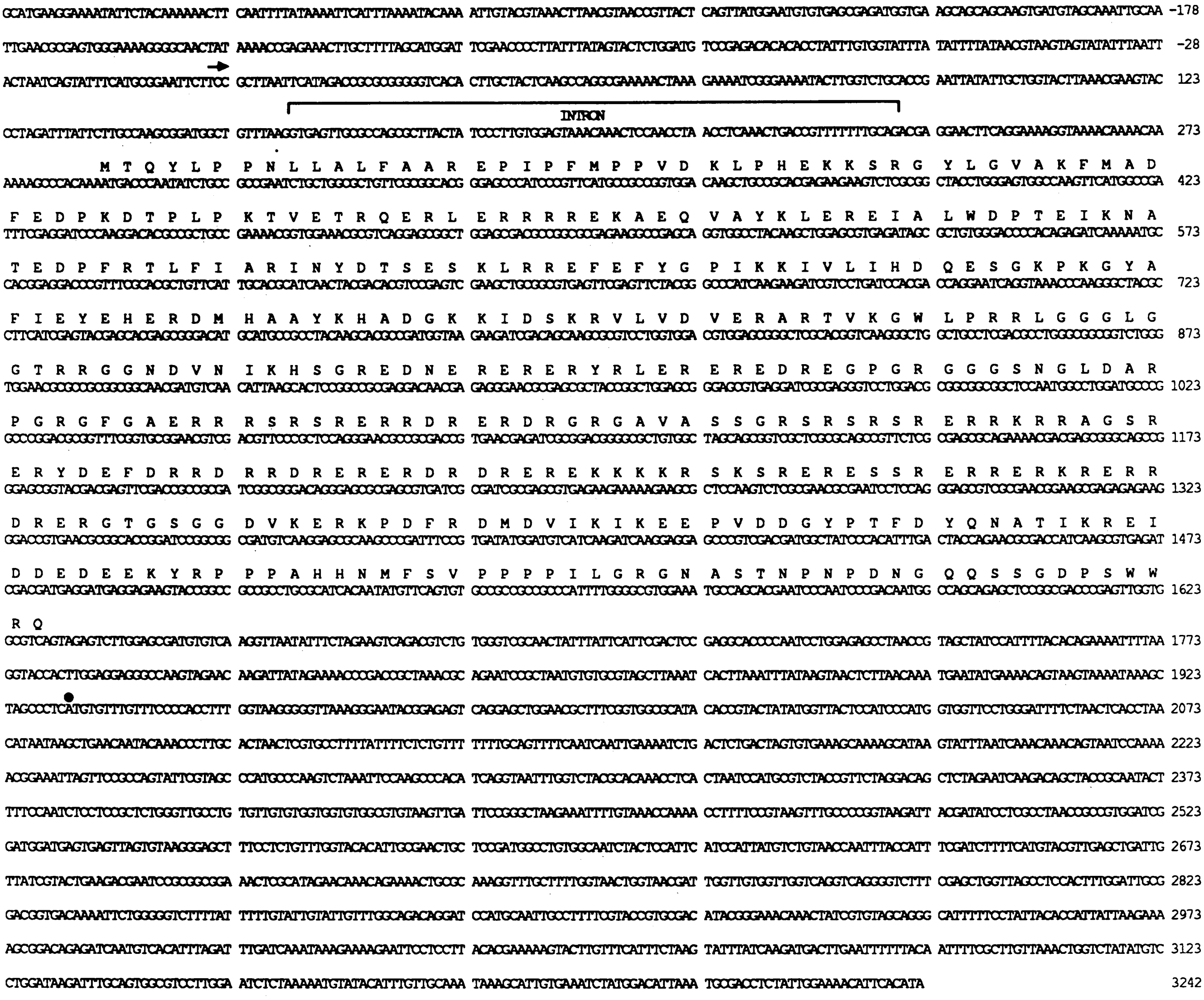
-178
-28
123
INTRON
273
M T Q Y L P P N L L A L F A A R E P I P F M P P V D K L P H E K K S R G Y L G V A K F M A D
423
F E D P K D T P L P K T V E T R Q E R L E R R R R E K A E Q V A Y K L E R E I A L W D P T E I K N A
573
T E D P F R T L F I A R I N Y D T S E S K L R R E F E F Y G P I K K I V L I H D Q E S G K P K G Y A
723
F I E Y E H E R D M H A A Y K H A D G K K I D S K R V L V D V E R A R T V K G W L P R R L G G G L G
873
G T R R G G N D V N I K H S G R E D N E R E R E R Y R L E R E R E D R E G P G R G G G S N G L D A R
1023
P G R G F G A E R R R S R S R E R R D R E R D R G R G A V A S S G R S R S R S R E R R K R R A G S R
1173
E R Y D E F D R R D R R D R E R E R D R D R E R E K K K K R S K S R E R E S S R E R R E R K R E R R
1323
D R E R G T G S G G D V K E R K P D F R D M D V I K I K E E P V D D G Y P T F D Y Q N A T I K R E I
1473
D D E D E E K Y R P P P A H H N M F S V P P P P I L G R G N A S T N P N P D N G Q Q S S G D P S W W
1623
R Q
1773
1923
2073
2223
2373
2523
2673
2823
2973
3123
3242

FIG. 5. Nucleotide sequence of the *Drosophila* U1 70K genomic and cDNA clones. Sequences from positions 327 to 435 and positions 2880 to 3242 are from λD331.1 insert genomic DNA. Sequences from positions 430 to 2885 are from cDNA clones pBS9.1 and λR3-2.1. See text for a full discussion of what sequences were determined in which clones. The horizontal arrow denotes a possible transcription start site at nucleotide 1 (see Discussion). The overline brackets the single 79-bp intron. The polyadenylation site for the 1.9-kb mRNA is indicated by the dot at position 1932. The derived U1 70K protein sequence is indicated above the nucleotide sequence, starting with the methionine at position 287. This sequence has been assigned the GenBank accession number M31162.

further. Assuming that the 1.4-kb cDNA clone is truncated at the 5′ end and that the two classes of RNA have identical or similar 5′ termini, an identical U1 70K protein would be encoded by the 3.1- and 1.9-kb mRNAs (see below).

Careful restriction mapping of the λD331.1 genomic clone and the λR3-2.1 and pBS9.1 cDNA clones revealed that there are no introns larger than the resolution of our agarose gels (about 20 bp for the small fragments used) within the two *Bam*HI fragments (nt 430 to 2885 of Fig. 5). We concluded that there are no introns present in this region. A single 79-bp intron was observed extending from position 161 to position 239 (designated by the overline in Fig. 5). The splice site sequences at both ends of this intron match their respective consensus sequences extremely well (34, 51), and two potential branch site consensus sequences (23, 38) are positioned 26 and 31 nt upstream of the 3′ end of the intron.

The sequence shown in Fig. 5 in the regions −327 to 435 and 2880 to 3242 was derived from genomic DNA, and the remaining sequence was derived from cDNA cloned inserts. Within those regions for which sequence was obtained from both genomic and cDNA clones (37 to 435, 645 to 965, and 2880 to 3242), three polymorphisms were observed. Two polymorphisms (positions 424 and 790, Fig. 5) were T-to-C changes in the cDNA sequence. Both of these were third-position changes that would not affect the predicted protein sequence. The third polymorphism (position 2932) was a C-to-T change in the 3′ untranslated portion of the cDNA sequence.

Predicted U1 70K protein sequence. The amino acid sequence of the *Drosophila* U1 snRNP 70K protein (Fig. 5 and 7A) was derived by conceptual translation of the *Drosophila* U1 70K cDNA. The start of translation is proposed to be at an ATG at position 287 which occurs in the context CAAAATGACC, a perfect match to the consensus sequence for *Drosophila* translational starts [(C/A)AA(A/C)**ATG**(a)(c)(c) (11)]. An ATG which occurs earlier in the sequence (at

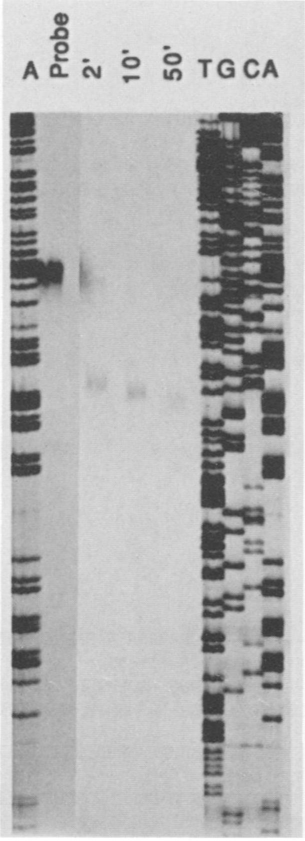

FIG. 6. S1 nuclease analysis of the downstream polyadenylation site. Nuclease S1 digests were done for the indicated times as described in Materials and Methods. The products were electrophoresed on a 6% polyacrylamide denaturing gel. Lanes: Probe, undigested 207-nt probe; 2', 24 μg of poly(A)$^+$ RNA digested for 2 min with 360 U of nuclease S1; 10', 24 μg of poly(A)$^+$ RNA digested for 10 min with 360 U of nuclease S1; 50', 24 μg of poly(A)$^+$ RNA digested for 50 min with 360 U of nuclease S1; T, G, C, and A, size markers; T, G, C, and A reactions derived from an unrelated clone of known sequence.

position 148) does not match this consensus, and initiation there would lead to the synthesis of only a tripeptide. Furthermore, the predicted amino acid sequence based on initiation at position 287 is extremely similar to those predicted from human and *Xenopus* cDNAs (see below), while there is no sequence similarity among conceptual translation products derived by using sequence 5' of this position in the three cDNAs (unpublished data). This starting position predicts a protein that contains 448 amino acids and that has a molecular weight of 52,879.

A comparison of the predicted sequences of the U1 snRNP 70K proteins from humans, *Xenopus laevis*, and *D. melanogaster* is shown in Fig. 7A. The U1 70K protein can be broken down into regions showing different patterns of conservation. The amino-terminal 214 amino acids are highly conserved among the three proteins, with 145 (68%) of these amino acids being identical in all three sequences. However, the pattern of conservation in the carboxy-terminal portion of the U1 70K protein is different. There is a distinct pattern of conserved arginine-rich regions in all three proteins. Although less conserved in sequence, the arginine-rich region has long stretches rich in arginine-glutamic acid, arginine-aspartic acid, and arginine-serine dipeptides in all three proteins. This difference is clearly shown in Fig. 7B, a matrix plot in which the strong identity between the *Drosophila* and human sequences in the amino-terminal portion of the protein is clearly depicted by a diagonal line which stands in clear contrast to the multiply repeated pattern of identity in the arginine-rich regions of the carboxy-terminal portion of the protein.

DISCUSSION

The central role of snRNPs in splicing (54) implies that an understanding of regulated splicing must depend on knowledge of the structure and function of snRNPs. Recently, several examples of regulated splicing have been described in *D. melanogaster* (see, for examples, references 2, 5, 16, 27, and 61), and it is likely that a combined genetic and biochemical approach will continue to make analysis of splicing in this organism particularly valuable. For these reasons, we undertook a study of the gene for the U1 snRNP 70K protein in *D. melanogaster*.

It has been noted previously that human autoimmune sera which recognize epitopes on U1 snRNP-specific proteins (43) will specifically immunoprecipitate U1 RNA from *Drosophila* cells (36). Thus, we expected that snRNP proteins in *D. melanogaster*, including U1 70K, would be highly conserved. This has proved to be true. Here, we identified a *Drosophila* gene which is clearly homologous to vertebrate genes encoding the U1 70K protein. To our knowledge, this is the first report of the cloning of a *Drosophila* gene for an snRNP protein. It appears from this one example that, like U snRNAs (reviewed by Guthrie and Patterson [19]), U snRNP proteins are highly conserved.

Expression of gene for *Drosophila* U1 70K protein. A number of results have indicated the possibility of developmental regulation of U1 70K gene expression in vertebrates. In *X. laevis*, a maternally deposited 2.0-kb U1 70K mRNA accumulates early in oogenesis and is stable for the remainder of oogenesis (15). Also, a 3.2-kb U1 70K mRNA is seen from the reactivation of U1 70K gene expression at the midblastula transition until late embryogenesis, and a 5.5-kb U1 70K mRNA is seen in swimming tadpoles. The structural differences among these developmentally regulated RNAs remain to be determined. Unlike *D. melanogaster*, the *Xenopus* coding region is interrupted by several introns and the *Xenopus* U1 70K protein is encoded by at least two genes. Thus, it is unclear what regulatory steps account for this pattern of expression of RNA products and what significance it might have for RNP structure. Similarly, in cultured human cell lines, an abundant U1 70K mRNA of 1.7 kb and a minor U1 70K mRNA of 3.9 kb are seen. Analysis of cDNAs indicates that there is at least some alternative splicing within the coding region which may give rise to different 70K proteins. In particular, there is an included or excluded exon that contains an in-frame termination codon and would give rise to a protein truncated 7 amino acids downstream of amino acid 159 (Fig. 7) if not removed (53).

We found no evidence of alternative splicing in the *Drosophila* gene. Although expression of the 3.1- and 1.9-kb transcripts appears to be developmentally regulated (Fig. 4), there is no difference between these two RNAs with regard to the expected protein product. The two mRNAs contain a single coding region uninterrupted by introns and differ only in their 3' untranslated regions. Furthermore, the single intron observed occurs within the 5' untranslated portion of the RNA. Finally, comparisons among the *Drosophila*, human, and *Xenopus* 5' untranslated sequences show little homology (15, 53, 59), which supports the positioning of the

```
A

N-terminus:

D.m.: MTQYLPPNLLALFAAREPIPFMPPVDKLPHEKK-SRGYLGVAKFMADFEDPKDTPLPKTVETRQERLERRRREKAEQVAYKLEREIALWDPTEIKNATEDPFR  102
      *** ********** * ***  **  ******     * * *     **** * * *   *** ** ** **** *      * *   ***    **  * *
H.s.: MTQFLPPNLLALFAPRDPIPYLPPLEKLPHEKHHNQPYCGIAPYIREFEDPRDAPPPTRAETREERMERKRREKIERRQQEVETELKMWDPHNDPNAQGDAFK  103
X.l.: MTQFLPPNLLALFAPRDPVPYLPPLDKLPHEKHHNQPYCGIAPYIREFEDPRDAPPPTRAETREERMERKRREKIERRQQDVENELKIWDPHNDQNAQGDAFK  103

RRM:

D.m.: TLFIARINYDTSESKLRREFEFYGPIKKIVLIHDQ-ESG--KPKGYAFIEYEHERDMHAAYKHADGKKIDSKRVLVDVERART  182
      *** ** **** ********* ***** *         *  ** ************** ***********  ******** **
H.s.: TLFVARVNYDTTESKLRREFEVYGPIKRIHMVYSK-RSG--KPRGYAFIEYEHERDMHSAYKHADGKKIDGRRVLVDVERGRT  183
X.l.: TLFVARVNYDTTESKLRREFEVYGPIKRIHIVYNKGSEGSGKPRGYAFIEYEHERDMHSAYKHADGKKIDGRRVLVDVERGRT  186

Glycine rich:

D.m.: VKGWLPRRLGGGLGGTRRGGNDVNIKHSGRED  214
      **** *************** **** **** *
H.s.: VKGWRPRRLGGGLGGTRRGGADVNIRHSGR-D  214
X.l.: VKGWRPRRLGGGLGGTRRGGADVNIRHSGR-D  217

Arginine rich:

D.m.: NERERERYRLEREREDREGPGRGGGSNGLDARPGRGFGAERRRSRSRERRDRERDRGRGAVASSG-------------------RSR-SRSRER-RKRRAGSRERYDEFDRR  305
                      *              *       *********** * * *       *                      * ****** *   * * **   *
H.s.: DTSRYDERPGPSPLPHRDRDRERERRERSRERDKE-RERRRSRSRDRRRRSRSRDKEERRRSRERSKDK---------DRDRKRRSSRSRERARRERER----KEELRGG  312
X.l.: DTSRYDER-----------DRERERDRRERSREREKEPRERRRSRSRERRRKSRSREKEERKRTREKSKDKDKEKDKDNKDRDRKRR-SRSRER-KRERDRDREKKEERV--  314

D.m.: -------------------------------------------------------------------DRRDRERERDRDREREKKKKRSKSRERESSRER-RERKRERRDRERGT  352
                                                                           **************    *  ***      * ***  * * **
H.s.: GGDMAEPSEAGDAPPDDGPPGELGPDGPDG---PEEKGR-----------------------------DRDRERRRSHRSERERRRDRDRDRDRDREHKR-GERGSE-RGRDEAR  393
X.l.: ---EAEVPEADDAPQDDAQIGDLGIDGIELKQEPEEKSRERDRERDRDREKGEKDRDKDRDRDRDRRRSHRDRDREKDRDRDRDRRRDRDRDRERDKDHKRERDRGDRSEKREERV  427

Carboxy terminus:

D.m.: GSGGDVKERKP-----DFRDMDVIKIKEEPVDDGYPTFDYQNATIKREIDDEDEEKYRPPPAHHNMFSVPPPPILGRGNASTNPNPDNGQQSSGDPSWWRQ  448
      * ** *          * ***           ***
H.s.: GGGGGQDNGLEGLGNDSRDMYMES-----EGGDGYLAPENGYLMEAAPE  437
X.l.: PDNGMVMEQAEE---TSQDMYLDQES-MQSGDGYLSTENGYMMEPPME  471
```

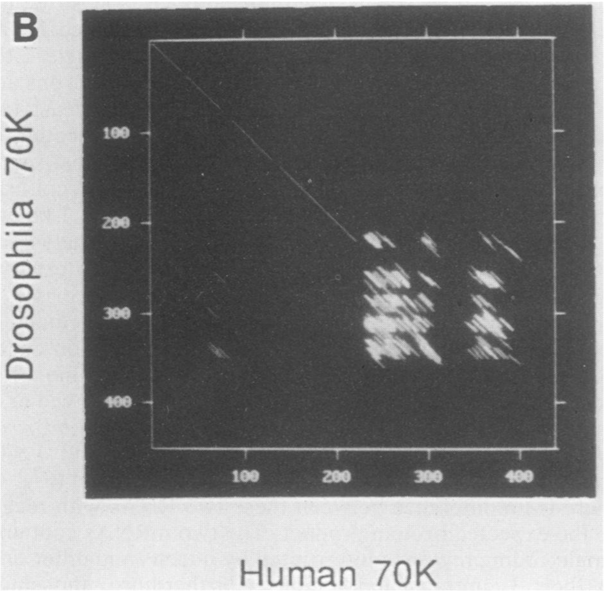

FIG. 7. Alignment of the human, *Xenopus*, and *Drosophila* U1 70K derived amino acid sequences. (A) The *Drosophila* U1 70K ORF (D.m.) aligned with the human FL 70K ORF (H.s.) from position 178 (nt 1212 in the FL 70K sequence [59]) and the *Xenopus* U1 70K (X.l.) (from positions 99 to 1512 in the Xc U1 70A cDNA clone [15]). The entirety of each sequence is shown, but the alignment is broken into the following regions for clarity. N-terminus, The N-terminal 102 amino acids in the *Drosophila* 70K protein. RRM, RNA recognition motif (46) is the region of the RNP consensus, originally described by Adam et al. (1) and conserved among RNA-binding proteins. Glycine-rich, The region between the RRM and the breakdown in alignment is rich in glycine residues. Arginine-rich, Regions rich in arginine (see text for a full discussion). Carboxy terminus, The C-terminal 96 amino acids in the *Drosophila* 70K protein. (B) The matrix plot of a comparison between *Drosophila* 70K protein and the human 70K protein. Regions of at least 50% amino acid identity including at least seven amino acids are denoted by a diagonal line. Note the contrast between the amino-terminal and carboxy-terminal portions.

initiating methionine determined in the human 70K proteins (45). In summary, the colinearity of cDNA and genomic sequences throughout the open reading frame (ORF) indicates that mRNAs with altered protein-coding potential are not produced from this gene.

U1 RNA recognition motif is highly conserved. Many proteins which are found associated with RNA, including poly(A)-binding proteins, snRNP proteins, and heterogeneous nuclear RNP proteins, share a common sequence

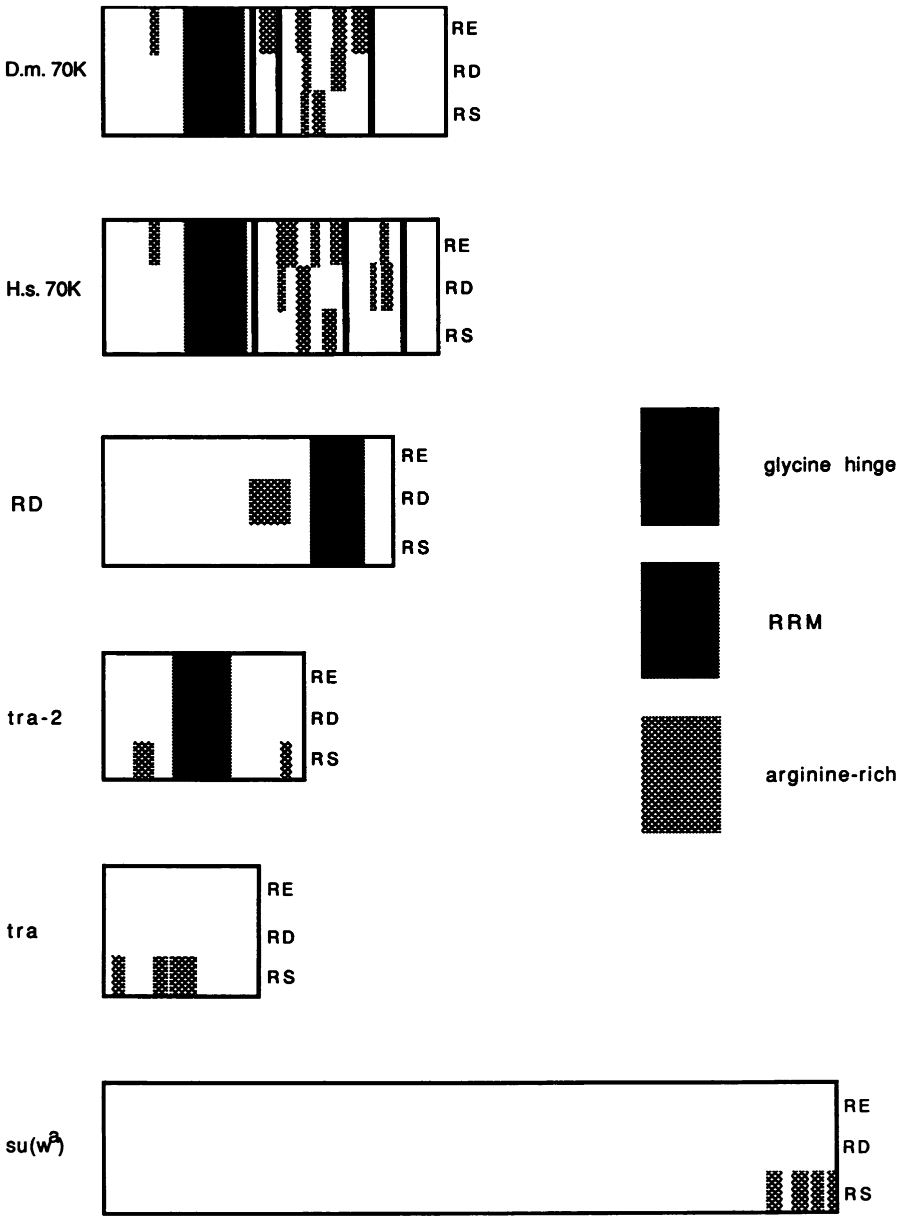

FIG. 8. Arrangement of sequence motifs in arginine-rich proteins. The derived amino acid sequences of the indicated proteins were searched by computer analysis for short stretches rich in arginine and glutamic acid, aspartic acid, or serine. A cluster was selected (denoted by a shaded box) whenever these 2 amino acids accounted for a minimum of 8 amino acids in a 10-residue stretch and both amino acids were present. Glycine hinge, Four of six residues are glycine. RRM, RNA recognition motif (46) is a conserved region encompassing the RNP consensus originally described by Adam et al. (1) that is most conserved among RNA-binding proteins. Arginine-rich, Regions rich in argininie that also contain aspartic acid, glutamic acid, or serine in at least 80% of the residues in a given stretch. D.m. 70K, *Drosophila* U1 70K protein (this report), H.s. 70K, the human FL 70K ORF from position 178 (nt 1212 in the FL 70K sequence [59]); RD, a mouse cDNA (29); *tra-2*, *Drosophila* transformer-2 protein (18); *tra*, *Drosophila* transformer protein (9); *su*(w^a), *Drosophila* suppressor-of-white-apricot protein (13). RE, RD, and RS refer to the dipeptides arginine-glutamic acid, arginine-aspartic acid, and arginine-serine, respectively.

element. First recognized as an 8-amino-acid consensus shared among yeast poly(A)-binding protein, several heterogeneous nuclear proteins, and the nucleolar protein nucleolin (57), this region of sequence similarity is now recognized as an 80- to 90-amino-acid stretch within which there are typically 15 to 20 amino acid identities when two unrelated family members are compared (see references 3 and 33 for recent reviews). Significantly, a slightly larger segment of the human U1 70K protein (amino acids 92 to 202 versus 104 to 183) has recently been shown to be capable of specific binding to U1 snRNA (46). We follow those authors in referring to this region as an RRM (RNA recognition motif). The coconservation of the first loop of U1 snRNAs (10 of 10 identity [36]) and the RRM of the 70K protein (28 of 30 amino acids in a region which includes the core consensus sequence [Fig. 7A]) strongly suggests that the binding interaction between U1 RNA and the U1 70K protein is identical in *D. melanogaster* and humans. Altogether, a relatively uniform 68% amino acid identity is observed in the first 214 amino acids. Note that this similarity extends from the initiating methionine, upstream of what has been identified as essential for binding to U1 RNA. In addition, this amino acid identity extends 12 amino acids downstream of what has been defined as minimal for specific binding to U1 RNA.

Conservation of Arg-Ser, Arg-Glu, and Arg-Asp motifs but not primary sequence or overall structure. The conceptually translated sequence from the 2.9-kb cDNA was compared with translated sequences from the GenBank library by using the program FASTA (42). Two classes of proteins were found: proteins with an RRM and proteins rich in arginine. Among the seven highest scores obtained are the heterogeneous nuclear RNP protein A1 (10), the rat helix-destabilizing protein (14), and hamster nucleolin (26), all proteins with previously identified RRMs (46). The most related gene in the data base (other than U1 70K genes from other species) is the *Drosophila* gene *tra-2*. In the absence of *tra-2* wild-type product, male-specific *dsx* mRNA is produced, leading to masculinization of chromosomally female flies (37). The similarity recognized by the computer was primarily within the RRM. However, *tra-2* also contains an arginine-rich region which contains alternating arginine and serine residues (18).

The second most related gene in the data base is the RD gene (29), a gene of unknown function which resides in the mouse major histocompatibility complex adjacent to the *Slp* gene, near the genes for the fourth component of complement and 21-hydroxylase. Although this gene has an RRM as well (R. Spritz, personal communication) (Fig. 8), the similarity recognized by the computer was stretches of alternating arginine and aspartic acid residues. Another protein that was identified as related by virtue of an arginine-rich region is the product of the *Drosophila transformer* gene. A mutation in this gene was identified in 1944 as transforming XX *Drosophila* flies from females to sterile males (55). More recently, it has also been shown to be required for the correct sex-specific splicing of the doublesex gene along with *tra-2* (37). Another gene which functions in splicing regulation in *D. melanogaster* and has an arginine-rich region is *suppressor-of-white-apricot* [*su*(w^a)] (13, 61). The *su*(w^a) gene has been shown to repress the splicing of its own mRNA (61). Bingham et al. (7) have suggested that these arginine-rich regions are diagnostic of a class of proteins that regulate RNA processing.

To examine and compare the arginine-rich regions of *tra-2*, *tra*, and *su*(w^a) proteins, we performed a computer search for short stretches rich in arginine and glutamic acid, aspartic acid, or serine. A cluster was selected whenever a minimum of 8 amino acids in a 10-residue stretch corresponded to the 2 amino acids in question and at least one of each corresponding amino acid was present. Using these criteria, it was clear that *tra-2*, *tra*, and *su*(w^a) contained exclusively arginine-serine stretches in their arginine-rich regions (Fig. 8). *tra-2* also has an RRM, which has led to the idea that this protein plays a direct role in splicing (18). Comparisons of these regulatory proteins with the *Drosophila* and human 70K proteins showed that the only common protein sequences among all the proteins are the arginine-serine stretches. Therefore, it is tempting to speculate that the arginine-serine stretches within the arginine-rich regions of these five proteins play a role in the regulation of splicing.

The *Drosophila* and human 70K proteins also have arginine-glutamic acid and arginine-aspartic acid stretches in addition to arginine-serine, and regions containing these tracts are separated by short glycine-rich regions in the 70K genes of both species. These glycine "hinges" are defined in Fig. 8 as stretches of amino acids where at least four of every six residues are glycine. As illustrated in Fig. 8, the entire arginine-rich region of the RD gene is composed solely of an arginine-aspartic acid stretch under the criteria set out above. It will be of interest to learn what the function of the RD gene is and whether it plays any role in splicing.

TABLE 1. Percentage of amino acids or dipeptides in arginine-rich regions

Protein	Region	%						
		R	E	D	S	ER + RE	DR + RD	SR + RS
Drosophila 70K	62–72	55	27	0	0	27	0	0
	212–232	38	38	18	0	52	5	0
	254–350	44	16	9	12	28	15	14
	394–402	11	44	33	0	11	0	0
Human 70K	63–73	45	36	0	0	36	0	0
	211–310	42	15	12	13	20	14	18
	344–393	42	18	16	6	16	30	4
RD	193–246	43	7	43	2	9	72	4
tra-2	39–59	38	5	0	48	0	0	38
	237–246	40	0	0	40	0	0	50
tra	13–26	50	14	7	29	21	7	43
	67–125	46	5	0	31	5	0	36
su(w^a)	850–871	32	5	5	45	5	0	18
	898–915	44	0	0	38	0	0	44
	923–933	45	0	0	36	0	0	45
	957–963	57	0	0	43	0	0	14

Table 1 summarizes the data of the computer search for single-amino-acid- and dipeptide-rich stretches. Regions in which at least 80% of the amino acids are R, E, D, or S are listed, along with their composition of RS, RD, or RE dipeptides. *tra-2*, *tra*, and *su*(w^a) have either all or almost all their dipeptides in RS tracts. For the RD gene, RD dipeptides compose almost all its arginine-rich region. Finally, for the *Drosophila* and human 70K proteins, most of the arginine-rich regions are made up of RE dipeptides; however, RD and RS dipeptides do account for a significant proportion of dipeptides in both proteins.

The occurrence of glycine-rich regions in the *Drosophila* and human 70K proteins may allow different regions within these proteins to function independently without imposing any steric hindrance. Once an snRNP has been assembled by specific interactions between the U1 snRNA and the RRM, the arginine-rich regions may be free to participate in other protein-RNA or protein-protein interactions. Although similar glycine-rich regions are not seen in the *Xenopus* 70K protein, other amino acids may play a similar role. Assuming that the arginine-rich region can function as an independent domain, what role might it have? It has been observed that light proteolysis destroys the ability of U1 snRNPs to bind 5′ splice site sequences in vitro (35). We think that the U1 70K arginine-rich region is an excellent candidate for the protein necessary for this binding. Additional possibilities include facilitation of U1 snRNP-U2 snRNP interactions or other aspects of spliceosome assembly and function.

What the data in Fig. 8 and Table 1 make clear is that the precise location of Arg-Ser, Arg-Asp, Arg-Glu, or Gly-rich segments is not conserved as well as is the overall composition within the carboxy-terminal region. Furthermore, it is clear that neither the RD gene nor the *Drosophila* splicing regulators have the variety of arginine-rich regions that the U1 70K protein does. Nevertheless, the presence of such similar regions in three of four genetically identified splicing regulators (the fourth being *Sex-lethal*, which has two RRMs but no arginine-rich segments [4]) makes it of considerable interest to determine the in vivo role played by these regions. To this end, we hope to use the cloned U1 70K gene to carry out an in vivo mutational analysis via P element-mediated transformation.

ACKNOWLEDGMENTS

We thank Charles Query and Jack D. Keene for providing the human U1 70K clone pSP64-β2. We thank Melissa Pepling for excellent technical assistance. We thank Tom Schwartz and Michael Levine for providing cDNA libraries, Mike Goldberg for providing a genomic library, Monica Driscoll for help and advice on double-stranded sequencing, and Molecular Dynamics for allowing us to use their densitometer at a product demonstration. We also thank Diane Robins and Jack D. Keene for critically reading the manuscript.

This work was supported by Public Health Service grant GM 37991 from the National Institute of General Medical Sciences, by a National Science Foundation Presidential Young Investigator award to S.M.M., and by Basil O'Conner Starter Scholar research award 5-630 from the March of Dimes Birth Defects Foundation.

LITERATURE CITED

1. **Adams, S. A., T. Nakagawa, M. S. Swanson, T. K. Woodruff, and G. Dreyfuss.** 1986. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. Mol. Cell. Biol. **6:**2932–2943.
2. **Baker, B. S.** 1989. Sex in flies: the splice of life. Nature (London) **340:**521–524.
3. **Bandziulis, R. J., M. S. Swanson, and G. Dreyfuss.** 1989. RNA-binding proteins as developmental regulators. Genes Dev. **3:**431–437.
4. **Bell, L. R., E. M. Maine, P. Schedl, and T. W. Cline.** 1988. *Sex-lethal*, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. Cell **55:**1037–1046.
5. **Bernstein, S. I., C. J. Hansen, K. D. Becker II, D. R. Wassenberg, E. S. Roche, J. J. Donady, and C. P. Emerson, Jr.** 1986. Alternative RNA splicing generates transcripts encoding a thorax-specific isoform of *Drosophila melanogaster* myosin heavy chain. Mol. Cell. Biol. **6:**2511–2519.
6. **Billings, P. B., R. W. Allen, F. C. Jensen, and S. O. Hoch.** 1982. Anti-RNP monoclonal antibodies derived from a mouse strain with lupus-like autoimmunity. J. Immunol. **128:**1176–1180.
7. **Bingham, P. M., T. Chou, I. Mims, and Z. Zachar.** 1988. On/off regulation of gene expression at the level of splicing. Trends Genet. **4:**134–138.
8. **Black, D. L., B. Chabot, and J. A. Steitz.** 1985. U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. Cell **42:**737–750.
9. **Boggs, R. T., P. Gregor, S. Idriss, J. M. Belote, and M. McKeown.** 1987. Regulation of sexual differentiation in D. melanogaster via alternative splicing of RNA from the *transformer* gene. Cell **50:**739–747.
10. **Buvoli, M., G. Biamonti, A. Ghetti, S. Riva, M. T. Bassi, and C. Horandi.** 1988. cDNA cloning of human hnRNP protein A1 reveals the existence of multiple mRNA isoforms. Nucleic Acids Res. **16:**3751–3770.
11. **Cavener, D. R.** 1987. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. Nucleic Acids Res. **15:**1353–1361.
12. **Chabot, B., and J. A. Steitz.** 1987. Multiple interactions between the splicing substrate and small nuclear ribonucleoproteins in spliceosomes. Mol. Cell. Biol. **7:**281–293.
13. **Chou, T.-B., Z. Zachar, and P. Bingham.** 1987. Developmental expression of a regulatory gene is programmed at the level of splicing. EMBO J. **6:**4095–4104.
14. **Cobianchi, F., D. N. Sen Gupta, B. Z. Zmudzka, and S. H. Wilson.** 1986. Structure of rodent helix-destabilizing protein revealed by cDNA cloning. J. Biol. Chem. **261:**3536–3543.
15. **Etzerodt, M., R. Vignali, G. Ciliberto, D. Scherly, I. W. Mattaj, and L. Philipson.** 1988. Structure and expression of a *Xenopus* gene encoding an snRNP protein (U1 70K). EMBO J. **7:** 4311–4321.
16. **Falkenthal, S., V. P. Parker, and N. Davidson.** 1985. Developmental variation in the splicing pattern of transcripts from the *Drosophila* gene encoding myosin alkali light chain result in different carboxyl-terminal amino acid sequences. Proc. Natl. Acad. Sci. USA **82:**449–453.
17. **Feinberg, A. P., and B. Vogelstein.** 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132:**6–13.
18. **Goralski, T. J., J.-E. Edström, and B. S. Baker.** 1989. The sex determination locus *transformer-2* of Drosophila encodes a polypeptide with similarity to RNA binding proteins. Cell **56:** 1011–1018.
19. **Guthrie, C., and B. Patterson.** 1988. Spliceosomal snRNAs. Annu. Rev. Genet. **22:**387–419.
20. **Hamm, J., M. Kazmaier, and I. Mattaj.** 1987. *In vitro* assembly of U1 snRNPs. EMBO J. **6:**3479–3485.
21. **Hinterberger, M., I. Pettersson, and J. A. Steitz.** 1983. Isolation of small nuclear ribonucleoprotein containing U1, U2, U4, U5 and U6 RNAs. J. Biol. Chem. **258:**2604–2613.
22. **Hultmark, D., R. Klemenz, and W. J. Gehring.** 1986. Translational and transcriptional control elements in the untranslated leader of the heat-shock gene *hsp22*. Cell **44:**429–438.
23. **Keller, E. B., and W. A. Noon.** 1984. Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. Proc. Natl. Acad. Sci. USA **81:**7417–7420.
24. **Krämer, A., W. Keller, B. Appel, and R. Lührmann.** 1984. The 5′ terminus of the RNA moiety of U1 small nuclear ribonucleoprotein particles is required for the splicing of messenger RNA precursors. Cell **38:**299–307.
25. **Lamond, A. I., M. M Konarska, and P. A. Sharp.** 1987. A mutational analysis of spliceosome assembly: evidence for splice site collaboration during spliceosome formation. Genes Dev. **1:**532–543.
26. **Lapeyre, B., H. Bourbon, and F. Amalric.** 1987. Nucleolin, the major nucleolar protein of growing eukaryotic cells: an unusual protein structure revealed by the nucleotide sequence. Proc. Natl. Acad. Sci. USA **84:**1472–1476.
27. **Laski, F. A., D. C. Rio, and G. M. Rubin.** 1986. Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. Cell **44:**7–19.
28. **Lerner, M., and J. A. Steitz.** 1979. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. Proc. Natl. Acad. Sci. USA **76:**5495–5499.
29. **Lévi-Strauss, M., M. C. Carroll, M. Steinmetz, and T. Meo.** 1988. A previously undetected MHC gene with an unusual periodic structure. Science **240:**201–204.
30. **Lührmann, R.** 1988. snRNP proteins, p. 71–99. *In* M. L. Birnstiel (ed.), Structure and function of major and minor small nuclear ribonucleoprotein particles. Springer-Verlag, New York.
31. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
32. **Maniatis, T., and R. Reed.** 1987. The role of small nuclear ribonucleoprotein particles in pre-RNA splicing. Nature (London) **325:**673–678.
33. **Mattaj, I. W.** 1989. A binding consensus: RNA-protein interactions in splicing, snRNPs, and sex. Cell **57:**1–3.
34. **Mount, S. M.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10:**459–472.
35. **Mount, S. M., I. Pettersson, M. Hinterberger, A. Karmas, and J. A. Steitz.** 1983. The U1 small nuclear RNA-protein complex selectively binds a 5′ splice site in vitro. Cell **33:**509–518.
36. **Mount, S. M., and J. A. Steitz.** 1981. Sequence of U1 RNA from Drosophila melanogaster: implications for U1 secondary struc-

ture and possible involvement in splicing. Nucleic Acids Res. **9:**6351–6368.

37. **Nagoshi, R. N., M. McKeown, K. C. Burtis, J. M. Belote, and B. S. Baker.** 1988. The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster. Cell **53:** 229–236.
38. **Nelson, K. K., and M. R. Green.** 1989. Mammalian U2 snRNP has a sequence-specific RNA-binding activity. Genes Dev. **3:**1562–1571.
39. **O'Hare, K., R. Levis, and G. M. Rubin.** 1983. Transcription of the *white* locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **80:**6917–6921.
40. **Padgett, R. A., S. M. Mount, J. A. Steitz, and P. A. Sharp.** 1983. Splicing of messenger RNA precursors is inhibited by antisera to small nuclear ribonucleoprotein. Cell **35:**101–107.
41. **Patton, J. R., and T. Pederson.** 1988. The *M*r 70,000 protein of the U1 small nuclear ribonucleoprotein particle binds to the 5′ stem-loop of U1 RNA and interacts with Sm domain proteins. Proc. Natl. Acad. Sci. USA **85:**747–751.
42. **Pearson, W. R., and D. J. Lippman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
43. **Pettersson, I., M. Hinterberger, T. Mimori, E. Gottlieb, and J. A. Steitz.** 1984. The structure of mammalian small nuclear ribonucleoproteins. J. Biol. Chem. **259:**5907–5914.
44. **Poole, S. J., L. M. Kauvar, B. Drees, and T. Kornberg.** 1985. The *engrailed* locus of Drosophila: structural analysis of an embryonic transcript. Cell **40:**37–43.
45. **Query, C. C., R. C. Bently, and J. D. Keene.** 1989. A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Cell **57:**89–101.
46. **Query, C. C., R. C. Bently, and J. D. Keene.** 1989. A specific 31-nucleotide domain of U1 RNA directly interacts with the 70K small nuclear ribonucleoprotein component. Mol. Cell. Biol. **9:**4872–4881.
47. **Query, C. C., and J. D. Keene.** 1987. A human autoimmune protein associated with U1 RNA contains a region of homology that is cross-reactive with retroviral p30^{gag} antigen. Cell **51:** 211–220.
48. **Ruby, S., and J. Abelson.** 1988. An early hierarchic role of U1 small nuclear ribonucleoprotein in spliceosome assembly. Science **242:**1028–1035.
49. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
50. **Seraphin, B., and M. Rosbash.** 1989. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell **59:**349–358.
51. **Shapiro, M. B., and P. Senapathy.** 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. **15:**7155–7174.
52. **Sharp, P.** 1987. Splicing of messenger RNA precursors. Science **235:**766–771.
53. **Spritz, R. A., K. Strunk, C. S. Surowy, S. O. Hoch, D. E. Barton, and U. Francke.** 1987. The human U1-70K snRNP protein: cDNA cloning, chromosomal localization, expression, alternative splicing and RNA-binding. Nucleic Acids Res. **15:** 10373–10391.
54. **Steitz, J. A., D. L. Black, V. Gerke, K. A. Parker, A. Krämer, D. Frendeway, and W. Keller.** 1988. Functions of the abundant U-snRNPs, p. 115–154. *In* M. L. Birnstiel (ed.), Structure and function of major and minor small nuclear ribonucleoprotein particles. Springer-Verlag, New York.
55. **Sturtevant, A. H.** 1945. A gene in *Drosophila melanogaster* that transforms females into males. Genetics **30:**297–299.
56. **Surowy, C. S., V. L. van Santen, S. M. Scheib-Wixted, and R. A. Spritz.** 1989. Direct, sequence-specific binding of the human U1-70K ribonucleoprotein antigen protein to loop I of U1 small nuclear RNA. Mol. Cell. Biol. **9:**4179–4186.
57. **Swanson, M. S., T. Y. Nakagawa, K. LeVan, and G. Dreyfuss.** 1987. Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins. Mol. Cell. Biol. **7:**1731–1739.
58. **Tabor, S., and C. C. Richardson.** 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc. Natl. Acad. Sci. USA **84:**4767–4771.
59. **Theissen, H., M. Etzerodt, R. Reuter, C. Schneider, F. Lottspeich, P. Argos, R. Lührmann, and L. Philipson.** 1986. Cloning of the human cDNA for the U1 RNA-associated 70K protein. EMBO J. **5:**3209–3217.
60. **White, P. J., P. B. Billings, and S. O. Hoch.** 1982. Assays for the Sm and RNP autoantigens: the requirement for RNA and influence of the tissue source. J. Immunol. **128:**2751–2756.
61. **Zachar, Z., T.-B. Chou, and P. M. Bingham.** 1987. Evidence that a regulatory gene autoregulates splicing of its transcript. EMBO J. **6:**4105–4111.
62. **Zhuang, Y., and A. M. Weiner.** 1986. A complementary base change in U1 snRNA suppresses a 5′ splice site mutation. Cell **46:**827–835.