

# The 64-Kilodalton Subunit of the CstF Polyadenylation Factor Binds to Pre-mRNAs Downstream of the Cleavage Site and Influences Cleavage Site Location

CLINTON C. MACDONALD,<sup>1</sup> JEFFREY WILUSZ,<sup>2</sup> AND THOMAS SHENK<sup>1\*</sup>

*Department of Molecular Biology, Howard Hughes Medical Institute, Princeton University, Princeton, New Jersey 08544-1014,<sup>1</sup> and Department of Microbiology and Molecular Genetics, UMDNJ—New Jersey Medical School, Newark, New Jersey 07103<sup>2</sup>*

Received 6 June 1994/Returned for modification 6 July 1994/Accepted 18 July 1994

**The CstF polyadenylation factor is a multisubunit complex required for efficient cleavage and polyadenylation of pre-mRNAs. Using an RNase H-mediated mapping technique, we show that the 64-kDa subunit of CstF can be photo cross-linked to pre-mRNAs at U-rich regions located downstream of the cleavage site of the simian virus 40 late and adenovirus L3 pre-mRNAs. This positional specificity of cross-linking is a consequence of CstF interaction with the polyadenylation complex, since the 64-kDa protein by itself is cross-linked at multiple positions on a pre-mRNA template. During polyadenylation, four consecutive U residues can substitute for the native downstream U-rich sequence on the simian virus 40 pre-mRNA, mediating efficient 64-kDa protein cross-linking at the downstream position. Furthermore, the position of the U stretch not only enables the 64-kDa polypeptide to be cross-linked to the pre-mRNA but also influences the site of cleavage. A search of the GenBank database revealed that a substantial portion of mammalian polyadenylation sites carried four or more consecutive U residues positioned so that they should function as sites for interaction with the 64-kDa protein downstream of the cleavage site. Our results indicate that the polyadenylation machinery physically spans the cleavage site, directing cleavage factors to a position located between the upstream AAUAAA motif, where the cleavage and polyadenylation specificity factor is thought to interact, and the downstream U-rich binding site for the 64-kDa subunit of CstF.**

Polyadenylation of precursor mRNAs involves a series of distinct steps: recognition of pre-mRNA substrates which contain the hexanucleotide sequence AAUAAA (19) by polyadenylation factors; endonucleolytic cleavage of the pre-mRNA at a distinct site five to 40 nucleotides (nt) 3' of the hexanucleotide; and addition of approximately 200 adenosine residues at the 3' end (11, 22, 29). In vivo, these processes are tightly coupled, but the steps can be readily uncoupled in vitro. Previous studies have identified U-rich or GU-rich regions, located as far as 50 nt downstream of the cleavage site, which influence polyadenylation both in vivo and in vitro (6, 13–15, 20, 21, 23, 24, 34). To date, the mechanism by which these downstream sequence elements influence polyadenylation has not been elucidated.

A 64-kDa (64K) polypeptide can be efficiently photo cross-linked to AAUAAA-containing RNA substrates that are undergoing polyadenylation in nuclear extracts of HeLa cells (31, 16). The 64K polypeptide is a subunit of cleavage and polyadenylation stimulation factor (CstF [27, 28]; also termed CF1 [7, 8]). A role for the factor in poly(A) addition was confirmed by the ability of antibodies specific for the 64K polypeptide to deplete CstF and block in vitro polyadenylation (27). The 64K polypeptide will specifically cross-link to AAUAAA-containing mRNAs as part of the CstF complex only in the presence of a second multisubunit factor (33) termed the cleavage and polyadenylation specificity factor (CPSF; formerly PF2, SF, or CPF [9, 16, 30]). CstF has been defined as a factor which, in the presence of CPSF, is required for the efficiency of polyadenylation (7, 8, 27, 28). Weiss et al. (30) have shown that CstF serves to stabilize a complex between CPSF and RNA with an

appropriate polyadenylation signal and downstream element. Binding of the 64K subunit of CstF to RNA is also influenced by the interaction of polyadenylation and splicing factors during recognition of 3'-terminal exons (18).

We have developed a procedure for mapping the site on an RNA molecule to which a polypeptide binds. Using this procedure, we have shown that the 64K subunit of CstF is photo cross-linked to the simian virus 40 (SV40) late and adenovirus type 2 L3 (Ad2L3) pre-mRNAs at U-rich regions that are located downstream of the site at which the pre-mRNA is cleaved during the polyadenylation reaction. Four consecutive U residues can substitute for the native U-rich element downstream of the cleavage site in the SV40 late pre-mRNA, and the site of cleavage on the SV40 pre-mRNA is influenced by the location of the stretch of U residues. These results lead us to propose (i) that the polyadenylation machinery physically spans the cleavage site, with CPSF interacting at the upstream AAUAAA motif and CstF interacting at a downstream U-rich sequence, and (ii) that CstF plays a role in cleavage site selection.

## MATERIALS AND METHODS

**RASH mapping.** Capped, <sup>32</sup>P-labeled RNAs were synthesized by using either SP6 or T7 RNA polymerase and [<sup>32</sup>P]UTP as described previously (31). RNase H synthetic hybrid (RASH) mapping was carried out as follows. A 275- $\mu$ l polyadenylation reaction mixture containing 188  $\mu$ l of HeLa nuclear extract, 72  $\mu$ l of 10% polyvinyl alcohol, 11  $\mu$ l of 200 mM phosphocreatine, 12.5 mM ATP, and ~450,000 cpm of SVL-3 (approximately 200 fmol) was distributed into 10 wells of a 96-well dish and incubated for 5 min at 30°C. Next, the mixture was irradiated with a germicidal light for 10 min at

\* Corresponding author. Phone: (609) 258-5992. Fax: (609) 258-1704. Electronic mail address: tshenk@molbiol.princeton.edu.

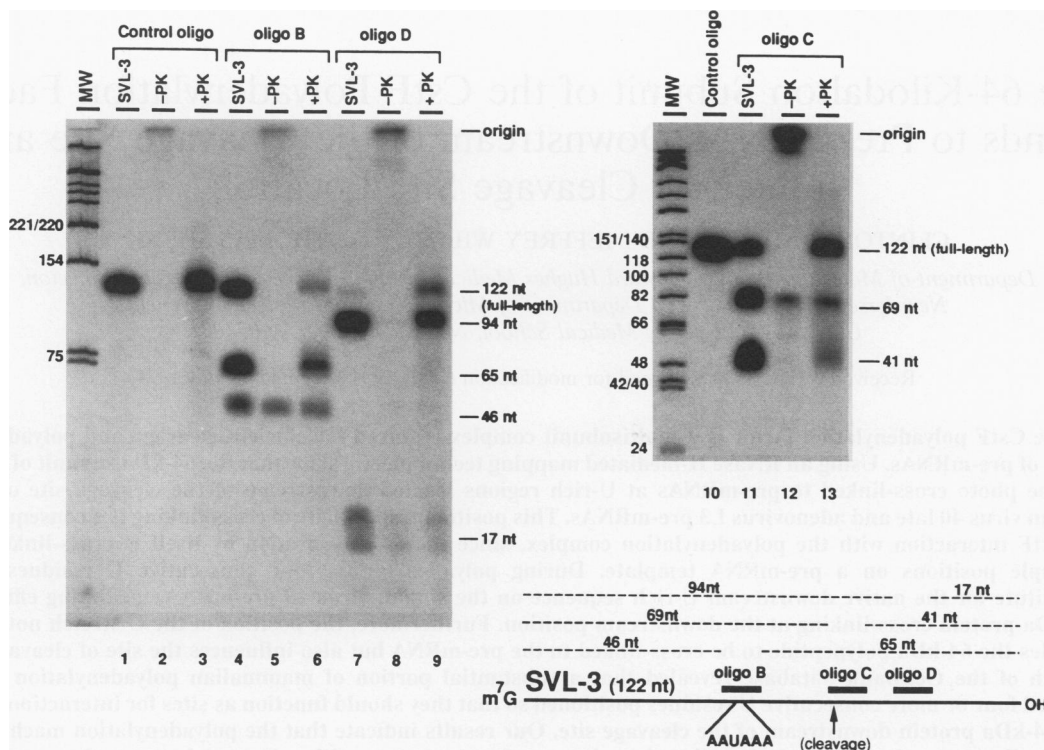


FIG. 1. The 64K CstF protein in nuclear polyadenylation extracts binds to a specific position downstream of the cleavage site on the SV40 late pre-mRNA. Position-specific binding to the SVL-3 substrate RNA by the 64K protein in nuclear extracts was mapped relative to the sites at which oligodeoxyribonucleotides (oligo) B, C, and D hybridize to the substrate. RASH mapping employed capped,  $^{32}\text{P}$ -labeled SVL-3 RNA. A 64K protein-specific monoclonal antibody was used to immunoprecipitate photo cross-linked protein-RNA complexes, after which oligodeoxyribonucleotides were hybridized, and the complexes were treated with RNase H. RNA products were analyzed by electrophoresis in a 6% polyacrylamide gel containing 8 M urea, and bands were visualized by autoradiography at  $-70^\circ\text{C}$  with an intensifying screen. MW, molecular weight markers; control oligo, a 15-nt oligonucleotide that does not hybridize to the substrate RNA; SVL-3, RNA which has been hybridized with oligonucleotide and treated with RNase H, included as a size marker; PK, proteinase K.

$4^\circ\text{C}$ . The material was recovered, mixed with  $750\ \mu\text{l}$  of radioimmunoprecipitation assay buffer (0.15 M NaCl, 50 mM Tris-HCl, 1% Nonidet P-40, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate [SDS], [pH 7.5]), boiled for 5 min, and immunoprecipitated with  $5\ \mu\text{l}$  of 13.5 mg of protein A-purified monoclonal antibody 3A7 per ml as described previously (27). Immunoprecipitated protein-RNA complex was eluted from protein-A Sepharose beads by two successive washes with  $200\ \mu\text{l}$  of NET-SDS (50 mM Tris-HCl, 0.3 M sodium acetate, 5 mM EDTA, 0.5% SDS) containing  $50\ \mu\text{g}$  of yeast tRNA per ml as a carrier for subsequent ethanol precipitation. Recovered protein-RNA complex (typically  $\sim 0.1\%$  of input radioactivity) was hybridized to oligodeoxyribonucleotides and hydrolyzed with RNase H as described previously (10). Hybridization and RNase H hydrolysis were at  $30^\circ\text{C}$  except for oligonucleotide C ( $25^\circ\text{C}$ ). Some samples were incubated with  $100\ \mu\text{g}$  of proteinase K in NET-SDS for 15 min at  $55^\circ\text{C}$  and then ethanol precipitated; other mixtures were precipitated with ethanol without proteinase treatment. Precipitates were resuspended in  $3\ \mu\text{l}$  of 80% formamide loading buffer, boiled for 5 min, and subjected to electrophoresis in 6% polyacrylamide gels containing 8 M urea.

**Bacterial expression of recombinant 64K protein.** Plasmid pHis64 was constructed from pHis64 $\Delta$ 247 (25) by replacing the *Hind*III fragment of that plasmid with a 1.2-kb *Hind*III-to-*Rsa*I fragment of pZ64-18 containing a DNA segment encoding the C-terminal 328 amino acids of the 64K protein. Expression of

this protein (termed rHis64) in *Escherichia coli* was exactly as described for rHis64 $\Delta$ 247 (25).

**Polyadenylation cleavage assays.** Pre-mRNA templates were pSVL, pSVL-3, pSP-AdL3, pSVL-GEM (16, 31, 32), pSVL-U15 (formerly termed pSVL-4U5 [32]), pSVL-U23, and pSVL-U36 (2). Plasmids pSVL-U23 and pSVL-U36 were constructed by first inserting the oligonucleotide 5'-TTTTTAA ACTGCAGTGCA-3' and its appropriate complement into the *Nsi*I site of pGEM-7Zf(-). The *Sst*I-to-*Pst*I and the *Bam*HI-to-*Pst*I fragments of the resulting plasmid were then inserted between the *Bsm*I-to-*Pst*I sites of pSVL to create pSVL-U23 and pSVL-U36, respectively. Polyadenylation and cleavage assays as well as UV cross-linking were done as described previously (31). Polyadenylation reaction mixtures contained 58% HeLa cell nuclear extract, 1 mM ATP, 16 mM phosphocreatine, 2.6% polyvinyl alcohol, and 10 fmol of capped  $^{32}\text{P}$ -labeled RNA; cleavage reactions were the same except that 1 mM AMP(CH<sub>2</sub>)PP and EDTA were substituted for the ATP and phosphocreatine.

## RESULTS

**RASH mapping of the binding site for the 64K polypeptide on pre-mRNAs.** We used the RASH mapping procedure to locate the site of 64K protein cross-linking to the SV40 late pre-mRNA. Capped,  $^{32}\text{P}$ -labeled pre-mRNA, corresponding to the 3'-terminal domain of the SV40 late transcription unit

(SVL-3), was prepared by *in vitro* transcription with SP6 polymerase. This substrate was added to a HeLa nuclear extract and incubated for a short time to allow polyadenylation complexes to form. Then the mixture was subjected to UV irradiation, and photo cross-linked 64K protein-RNA complexes were isolated by immunoprecipitation with a 64K protein-specific monoclonal antibody (27). The protein-RNA complex was hybridized to an oligodeoxyribonucleotide complementary to 12 to 15 nt of the substrate RNA, and the resulting RNA-DNA hybrid was hydrolyzed with RNase H (10). Finally, the cleaved RNA was analyzed by denaturing gel electrophoresis to determine which of the resulting fragments contained the cross-linked 64K protein. RNA-protein complexes were readily distinguished from free RNA because the complexes failed to enter the gel during electrophoresis.

Capped, 122-nt SVL-3 RNA was not hydrolyzed by RNase H when it was incubated with a noncomplementary oligodeoxyribonucleotide (Fig. 1, lane 1). When the SVL-3 RNA was photo cross-linked to the 64K protein in a polyadenylation reaction and immunoprecipitated with a 64K protein-specific monoclonal antibody, it failed to enter the polyacrylamide gel during electrophoresis (Fig. 1, lane 2). Digestion with proteinase K allowed the RNA to enter the gel (Fig. 1, lane 3), confirming that the abnormal migration of the complex was due to the cross-linked protein. The inability of RNA molecules containing cross-linked protein to enter the gel during electrophoresis provided a simple method for localization of 64K cross-linking sites to specific RNA fragments. Further, it was clear that the results would not be obscured by contaminating RNA lacking bound 64K protein, since no free, normally migrating RNA was evident in the immunoprecipitated sample (Fig. 1, lane 2). Finally, the reaction was specific for the 64K protein, since control experiments using a variety of nonspecific antibodies failed to precipitate an RNA-protein complex (data not shown).

Next, free SVL-3 RNA was hybridized to oligonucleotide B, which was complementary to nt -24 to -13 relative to the cleavage site utilized during polyadenylation (Fig. 1, diagram). RNase H hydrolysis produced three bands (Fig. 1, lane 4); the 122-nt species corresponded to unhydrolyzed SVL-3 RNA, while the 65- and 46-nt bands represented sequences 3' and 5' to the oligonucleotide B sequence, respectively. When immunoprecipitated SVL-3 RNA-protein complex was cleaved by using oligonucleotide B and RNase H, the 122-nt full-length and the 65-nt product were retained at the origin (Fig. 1, lane 5). Both RNAs could be released by proteinase K digestion (Fig. 1, lane 6), indicating that the 64K protein is cross-linked within the 65-nt fragment. Similar analyses with oligonucleotide D (complementary to nt +27 to +38; Fig. 1, lanes 7 to 9) and oligonucleotide C (complementary to nt +2 to +13; Fig. 1, lanes 11 to 13) led to the conclusion that the 64K CstF protein is cross-linked within a 13-nt region located between +14 and +26 relative to the site at which the SVL pre-mRNA is cleaved during polyadenylation (see Fig. 5A).

We also used the RASH technique to map the site of 64K protein cross-linking to a second polyadenylation substrate, the Ad2L3 pre-mRNA. As expected, incubation with a nonspecific oligodeoxyribonucleotide failed to induce hydrolysis of the 164-nt Ad2L3 RNA with RNase H (Fig. 2, lane 1). Free Ad2L3 RNA was incubated with oligonucleotide E (complementary to nt +4 to +15 relative to the cleavage site; Fig. 2, diagram). Hydrolysis with RNase H produced three bands: a 164-nt species corresponding to uncleaved RNA and 110- and 42-nt bands representing RNA segments 5' and 3', respectively, to the oligonucleotide E sequence. When immunoprecipitated Ad2L3 RNA-protein complex was cleaved by using oligonu-

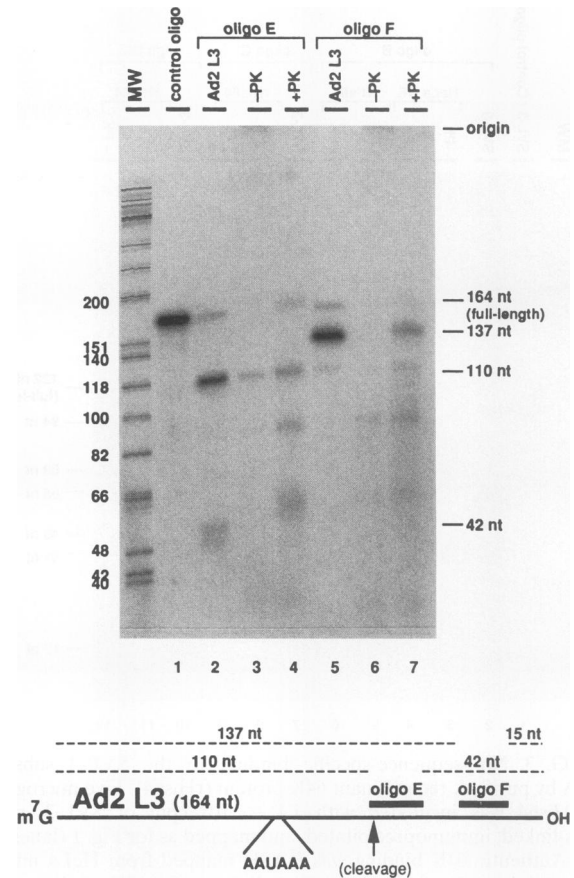


FIG. 2. The 64K CstF protein in nuclear polyadenylation extracts binds to a specific position downstream of the cleavage site on the Ad2L3 pre-mRNA. Capped,  $^{32}\text{P}$ -labeled Ad2L3 RNA was produced by runoff transcription with the SP6 RNA polymerase. RASH mapping was carried out as described with oligodeoxyribonucleotides (oligo) complementary to Ad2L3 RNA. The positions to which the oligonucleotide hybridize to the RNA substrate are diagrammed at the bottom. Note that the gel system used was unable to resolve the 15-nt product of RNase H hydrolysis by oligonucleotide F. MW, molecular weight markers; control oligo, a 15-nt oligonucleotide that does not hybridize to the substrate RNA; Ad2L3, Ad2L3 RNA which has been hybridized with oligonucleotide and treated with RNase H, included as a size marker; PK, proteinase K.

cleotide E and RNase H, the 42-nt segment was retained at the origin, while the 110-nt segment migrated normally during electrophoresis (Fig. 2, lane 3). On treatment with proteinase K, the 42-nt fragment generated from the RNA-protein complex entered the gel (Fig. 2, lane 4). Therefore, the 64K polypeptide binds within the 42-nt fragment, downstream of oligonucleotide E. A similar analysis placed the 64K binding site upstream of oligonucleotide F (Fig. 2, lanes 5 to 7). Thus, the 64K protein is cross-linked between nt +16 and +30 relative to the cleavage site on the Ad2L3 pre-mRNA (see Fig. 5A).

**Specific binding by the 64K protein requires associated polyadenylation factors.** The 64K polypeptide was expressed in *E. coli* as a fusion protein with six histidine residues at its amino terminus (rHis64) and purified by nickel chelate chromatography. This fusion protein was subjected to the RASH mapping procedure to determine whether it had the intrinsic ability to bind specifically to downstream U-rich sites on this

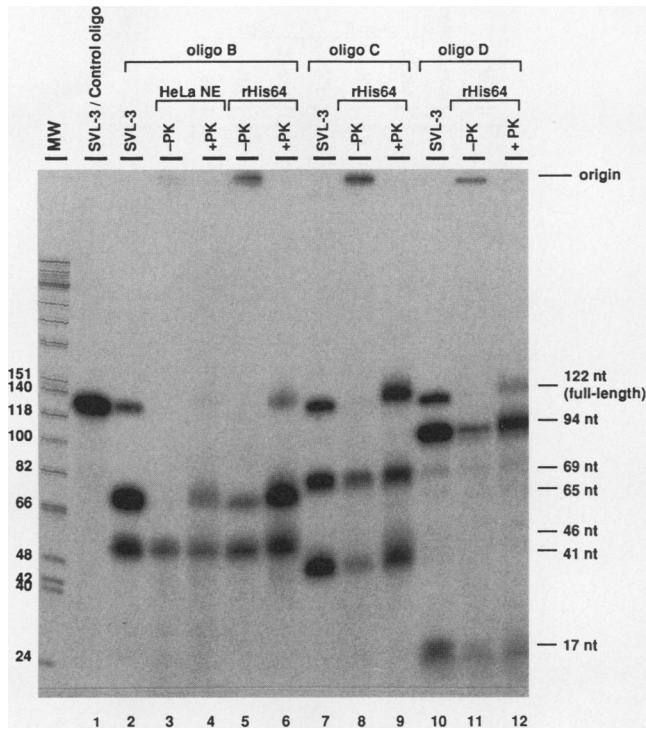


FIG. 3. Non-sequence-specific binding to the SVL-3 substrate RNA by purified, recombinant 64K protein (rHis64). Five micrograms of rHis64 was incubated with  $1.1 \times 10^6$  cpm of SVL-3, photo cross-linked, immunoprecipitated, and mapped as for Fig. 1 (lanes 5 to 12). Authentic 64K binding was RASH mapped from HeLa nuclear extract, using oligonucleotide (oligo) B (lanes 2 and 3). MW, molecular weight markers; control oligo, a 15-nt oligonucleotide that does not hybridize to the substrate RNA; PK, proteinase K.

pre-mRNA. Recombinant 64K protein was previously shown to photo cross-link to RNA, but in a non-AAUAAA-dependent manner (reference 25 and data not shown). The interaction of purified rHis64 protein with the SVL-3 pre-mRNA was mapped by the RASH procedure (Fig. 3). SVL-3 RNA was photo cross-linked to bacterially expressed rHis64, immunoprecipitated with antibody to the 64K protein, incubated with oligonucleotide B and RNase H, and analyzed by electrophoresis on a denaturing polyacrylamide gel. Reduced levels of RNA corresponding to both the 46- and 65-nt RNA products are evident in the gel (Fig. 3, lane 5), but no RNA band corresponding to the full-length 122-nt species was present. Incubation with proteinase K resulted in an increase of radioactivity in the 46- and 65-nt bands, as well as the reappearance of a band at 122 nt (Fig. 3, lane 6). This is the result one would expect if the 64K protein was cross-linked to SVL-3 RNA at positions both upstream and downstream of oligonucleotide B. For comparison, the specific site to which 64K protein within a HeLa nuclear extract binds to SVL-3 RNA was assayed by using oligonucleotide B (Fig. 3, lanes 3 and 4). It bound downstream but not upstream of oligonucleotide B, as before. RASH mapping of rHis64 on SVL-3 with oligonucleotides C and D yielded similar results; i.e., the protein appeared to bind both upstream and downstream of the sites to which the oligonucleotides bound (Fig. 3, lanes 7 to 12). Thus, it appears that isolated 64K protein can bind to the SVL-3 pre-mRNA, but it binds at multiple sites.

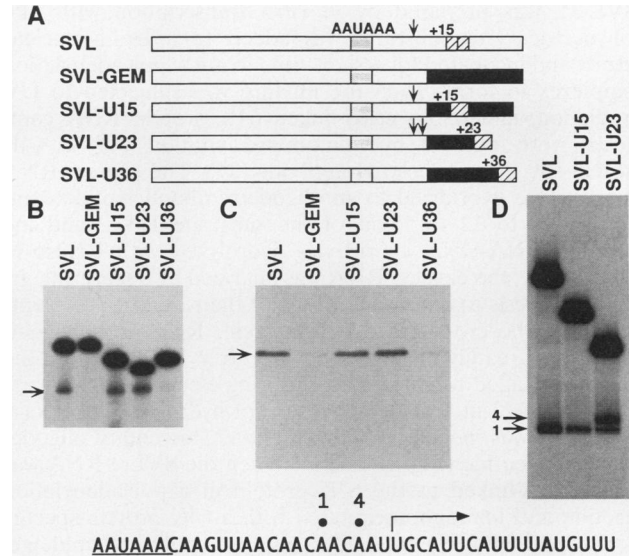


FIG. 4. Five consecutive U residues can serve as a binding site for the 64K CstF protein in nuclear polyadenylation extracts and influence both processing efficiency and the position of the cleavage site. (A) Diagram of the SVL pre-mRNA and variant substrate RNAs that contain substituted downstream sequences. The AAUAAA motif is designated by a stippled bar; the substituted domain, derived from pGEM, is designated by a solid bar; and the position of naturally occurring U-rich sequences in SVL or the sites of inserted runs of five U residues are designated by cross-hatched bars. Arrows mark cleavage sites. (B) Assay for cleavage of substrate RNAs in HeLa cell nuclear extracts. EDTA and AMP(CH<sub>2</sub>)PP were added to block addition of a poly(A) stretch to cleaved RNAs. The arrow marks the position of the cleavage products. (C) Assay for photo cross-linking of the 64K CstF protein to substrate RNAs in HeLa cell nuclear extracts. Polyadenylation reactions were photo cross-linked with a germicidal light and then treated with RNase A. The 64K protein was immunoprecipitated with a monoclonal antibody, and protein containing <sup>32</sup>P-labeled RNA oligomers was analyzed by SDS-polyacrylamide gel electrophoresis and autoradiography. The arrow marks the position of the 64K protein. (D) Analysis of polyadenylation cleavage products in a 5% polyacrylamide sequencing gel containing 8 M urea. Bands corresponding to cleavage at positions +1 and +4 are marked by arrows, and the cleavage sites are shown in the SVL sequence at the bottom. The dots above the sequence represent the sites of cleavage, and the arrow represents sequences which were replaced in SVL-U15, -U23, or -U36.

**Four consecutive U residues can substitute for the native 64K protein binding site in the SVL pre-mRNA.** The downstream binding site for the 64K polypeptide on the SVL and Ad2L3 pre-mRNAs is U rich. We previously reported that a stretch of five U residues can functionally substitute for the native downstream domain of both the SV40 late and SV40 early pre-mRNAs to activate polyadenylation (32). Therefore, a series of SVL derivatives were analyzed to test the ability of the 64K protein to interact specifically with a short stretch of U residues downstream of the cleavage site (Fig. 4A). SVL-GEM contains an unrelated polylinker sequence with no more than two consecutive U residues in place of the native SVL downstream sequences beginning at nt +6. Additional variants similar to SVL-GEM were generated, except that they contained five consecutive U residues at +15, +23, or +36 relative to the cleavage site. These substrates were tested for their abilities to undergo cleavage (Fig. 4B) and to cross-link to the 64K CstF protein (Fig. 4C). The wild-type SVL pre-mRNA



TABLE 1. Mammalian polyadenylation sites containing U-rich elements

Gene <sup>a</sup>	GenBank accession no.	Sequence		Position <sup>b</sup>
		Upstream <sup>c</sup>	Downstream <sup>d</sup>	
Rnnadh24	X14724	<u>AAATAAA</u> TATGAATCTCTATCTACCT	AAAT <u>TTTTT</u> TATACTACTTGTTCATTTCCATC	4
Humcol18ax	L22548	<u>AAATAAA</u> AGGAAGCCAAAGAGT	GTAT <u>TTTTT</u> TTAAAAGTTTAAAACAGAAGCCT	4
Hss31g	Z11166	<u>AAATAAA</u> CATTATCATACAAAAAC	AGAAA <u>TTTTT</u> GTAGAAGTAATAAAGATGAA	6
Rnlapg	X54626	<u>AAATAAA</u> CTTTTTTGGAGAATGTTT	AAAAAC <u>TTTTT</u> GGGGGCAGTAGTTGGCTTTT	8
Bovcasb2	M21343	<u>AAATAAA</u> TTAATCCTTAGGCAT	ATTCAGT <u>TTTTT</u> CTGTCTTTATTATATTTTT	8
Ratcrp2a	M84011	<u>AAATAAA</u> CTTTTTTGGAGAATGTTT	AAAAAGC <u>TTTTT</u> GGGGGCAGTAGTTGGCTTT	9
Musahr11	L19756	<u>AAATAAA</u> CTAACCACCAAAAT	AAATAGT <u>TTTTT</u> CATGCAAGTTTTTCTTTCT	9
Humant1	J04982	<u>AAATAAA</u> GACCACTTAAATGCAC	GCTTCTAT <u>TTTTT</u> TGAACCTTATTAACCTG	9
Hummanr30	M93221	<u>AAATAAA</u> CATTAAAGACAAGGCTATTTTTAAAT	AGATTTGCAT <u>TTTTT</u> GGTGATCTTAGTATAAAT	10
Walblgx07	L14960	<u>AAATAAA</u> CTGAAAATCATGGATACC	ATTTGATTC <u>TTTTT</u> GGCTTTCTCTTTTTTATA	10
Mmn10g	X16995	<u>AAATAAA</u> AAATAGACATGTAGTTG	AAATTGAGT <u>TTTTT</u> CTGT <u>TTTTT</u> CTGT <u>TTTTT</u>	11
Hsacpp	X52174	<u>AAATAAA</u> GTCACTTATAGGAAGTTC	AAAAGCAT <u>TTTTT</u> GTATTTGAACCTTATTAACCTG	12
Humembpa	M34462	<u>AAATAAA</u> ATGGGTTTTACTG	AAATGGATTAT <u>TTTTT</u> CTCCTCTGATCGCGGA	12
Hscampsg	X13461	<u>AAATAAA</u> TGAGATATTCGTCCCTT	ATTTTCA <u>TTTTT</u> CTTTTTTTTTTTGGTTTCCAT	12
Ratbp	M14050	<u>AAATAAA</u> TGTTTGTATTTT	ACACTGGTCTAG <u>TTTTT</u> GTGTAGTCT	13
Bovmarcks	M24638	<u>AAATAAA</u> GTGTAATTATGTATT	ACCAGCAGGGT <u>TTTTT</u> TAACCTGATGACTATTG	13
Humkcs	D10522	<u>AAATAAA</u> GTGTAATTATGTATT	ACCAGCAGGGT <u>TTTTT</u> TAACCTGATGACTATTG	13
Hsafpa4	Y00264	<u>AAATAAA</u> ATAAC	CCCGGGCAAGACT <u>TTTTT</u> CTTTGAAGGATGACT	13
Humodca2	M34158	<u>AAATAAA</u> GTATCTTGAAATAAAT	AGGCATGGGACG <u>TTTTT</u> TATGGTGTGTTCAT	14
Hummet2	J00271	<u>AAATAAA</u> ACAGCTTTGACTTG	ATTCGTCTCTG <u>TTTTT</u> CTTTTAACTGATGCTTTA	14
Hsapoiaia	X01038	<u>AAATAAA</u> CGTTTCCAAAGTGGG	AAGCAGCTTCTTCT <u>TTTTT</u> GGGAGAATAGAGG	15
Humpcra	M11228	<u>AAATAAA</u> AAAACAACACACG	AAGCCACTAGAGC <u>TTTTT</u> CTCCAGGGCTTTGGG	15
Mushox111a	M93148	<u>AAATAAA</u> TTCTCTTTTGCTAT	AAGTAGCTTCTTAT <u>TTTTT</u> CTTCCCCTCT	15
Ratpam27	L01693	<u>AAATAAA</u> GTGTTTTCAGAGCATT	AGCAAGTCAGTGA <u>TTTTT</u> CTTTTATGCTTTGG	15
Rnu02316	U02316	<u>AAATAAA</u> GTATTTACCTT	AAATTAACAATTTA <u>TTTTT</u> TATTTAGCAGTT	16
Humfcxii3	M17466	<u>AAATAAA</u> GTGCTTTGAAAATGC	TGAGAAGGAAAGCT <u>TTTTT</u> CTTTCATGGGTCC	16
Hsjepr	X60001	<u>AAATAAA</u> TATATTTTTGTAC	AAAACCTGACTTCCAGT <u>TTTTT</u> CTTGAAGGAA	17
Hslipa10	X75497	<u>AAATAAA</u> ATTGAGTGAGACTTATAGGCC	AAAGAATTGTGTGATG <u>TTTTT</u> GTGTG	18
Mmcrelm	X15842	<u>AAATAAA</u> CTCATATTATGTTGAGGCCATAGGC	AAAGTGACCCCTAAGGGT <u>TTTTT</u> CTGAGGTTTCT	18
Hummrp8a	M21005	<u>AAATAAA</u> GTCAATCAATACCTC	ATGCCTCTCTCTTATGCT <u>TTTTT</u> GTGGAAATGAG	18
Humtpa	K03021	<u>AAATAAA</u> TTTCAGATATATTTTTTCACAGTTTTTCC	AAAATCAGAGTGGAAATGGT <u>TTTTT</u> GTATAGAT	19
Humcbp	M15353	<u>AAATAAA</u> ATGTGTACAGTGAATGTCTTTT	AGCAACTAGATTGTGCTCTTTATTTTACTCCA	19
Humsp4040a	L00974	<u>AAATAAA</u> ACTGTCTTGTGAGCTG	ATCGCTGGAGGGTCTCT <u>TTTTT</u> TATGTTGAG	19
Mmb3a	X72862	<u>AAATAAA</u> TATTATGAAATGTG	GTTTTGCCCTGCTCGCCT <u>TTTTT</u> CTCTGGGTT	19
Muspkcd10	M19960	<u>AAATAAA</u> AGTTTTGTTTT	AAAATAAAAGTGTGCTG <u>TTTTT</u> CTCCGACTT	20
Humtnp1	M59924	<u>AAATAAA</u> ACTTTGAAAGC	AGAAATGTTTACTTTGATCT <u>TTTTT</u> GGTTGGAA	20
Humaldb2	M15656	<u>AAATAAA</u> TATTTCTATAAACCTTC	AGCTGCTATCGGGTTTCACT <u>TTTTT</u> CTGCTCTT	20
S45011	S45011	<u>AAATAAA</u> AGCTAAATTTTGG	ATCTGGTGACTGGTTTGAAC <u>TTTTT</u> TTTCACT	21
Hshlaeut3	X64881	<u>AAATAAA</u> AGCTGTGTTGAGCATC	ATTTGATTTTCAATTTGTGCGT <u>TTTTT</u> GTGCCTT	21
Humcypii	J02843	<u>AAATAAA</u> TATTTTCCCAGAAATAT	AAATAAATCATCACATGATTTATTTTAACTAT	22
Ratkalp15	M62357	<u>AAATAAA</u> AGATGTTCTACTTG	AACTTGATGGGACAGAGCACCT <u>TTTTT</u> CATATA	22
Humpregspe	L17043	<u>AAATAAA</u> ACATGTACCACATTTGCT	AATACTGTTCTCTTAAAATAATTTTAAAAGA	22
Hsfol	X69516	<u>AAATAAA</u> CAGACTGTTTCTAAT	AATTCATGCTGTGGAATTTGTTTGGTTGT	22
Hsi105	X00695	<u>AAATAAA</u> TTTGATAAATAT	AAACAAGCCTGGATATTTGTTATTTTGGAAA	23
Ocace3p	X62549	<u>AAATAAA</u> AGCCCTTTATACCAATATCAAGAGAT	AAGCACTGTTGACAGTTTGGTGT <u>TTTTT</u> ATCTT	23
Hummk	D10604	<u>AAATAAA</u> AGCTCTCTTTTTTAAATAT	ATAAAAGCCCTTCCCAAGGAGTTTGTCTGT	23
Muslipol12	M69260	<u>AAATAAA</u> ACAATTTTATAAGAC	AATTTGGTGTGCATTTGCTCTCTCT <u>TTTTT</u> GTGA	24
Hum2c9x07	L16883	<u>AAATAAA</u> AGCATTATTTATTTGCTG	AGTCAGTTTATGACCTTCTCTCT <u>TTTTT</u> TATG	25

<sup>a</sup> Gene designation as reported in GenBank entry. Poly(A) addition sites were identified by searching 52,214 nonviral, mammalian sequences for eight or more A residues between 5 and 40 nt downstream of the canonical AATAAA (1–3). Duplicates and sequences such as poly(A/T) were eliminated to yield 1,856 sequences. The database was again searched for these sequences but lacking the 3' terminal poly(A) residues. This search identified 180 gene segments encoding sequences upstream and downstream of the site of poly(A) addition.

<sup>b</sup> Position of downstream T-stretch from cleavage site.

<sup>c</sup> From the AATAAA motif to the inferred site of cleavage.

<sup>d</sup> Downstream of the inferred site of cleavage.

both CPSF and CstF are required for efficient cross-linking of the 64K protein to pre-mRNA (33).

Four consecutive U residues are sufficient to direct specific binding of the 64K polypeptide and support cleavage of the SVL polyadenylation substrate (Fig. 4 and reference 3). The site mapped in the SVL downstream domain contains four consecutive U residues, but the Ad2L3 U-rich downstream domain does not (Fig. 5A). This finding suggests that there is flexibility or degeneracy in the composition of natural U-rich

elements. Our analysis demonstrates that this U-stretch can activate polyadenylation of an SVL derivative lacking its natural downstream domain when it is located between 15 and 23 nt downstream from the site of cleavage (Fig. 4D), and study of additional SVL derivatives has refined the permissible location to the segment between 6 and 25 nt downstream from the cleavage site (3). Movement of the U stretch within the downstream domain can alter the choice of cleavage site (Fig. 4). Earlier studies of the *Xenopus*  $\beta$ -globin gene (11) and the



SV40 late gene (5) found that deletions downstream of the polyadenylation signal can affect the position of the cleavage site. These observations can now be interpreted to result from displacements caused by the deletions of downstream 64K protein binding sites relative to the AAUAAA motif in pre-mRNAs. The importance of the location of the U-rich downstream element is further underscored by recent analysis of transcripts encoded by type I human T-cell leukemia virus (HTLV-I) (1). Secondary structures within the Rex response element of HTLV-I function to precisely position the downstream element relative to the AAUAAA motif and allow polyadenylation. Disruption of the Rex response element secondary structure positions the downstream motifs too far from the AAUAAA signal and prevents polyadenylation. The HTLV-I downstream element contains a GU-rich motif adjacent to a stretch of four U residues.

If downstream U-rich elements are important for polyadenylation site function, one would expect to find such elements in appropriate regions downstream of many polyadenylation sites. We tested this hypothesis by searching the GenBank database (version 81.0, February 1994). We were able to identify a total of 180 nonviral, mammalian polyadenylation sites for which the sequence downstream of the cleavage site was available. A survey of these sites revealed that 48 contain a stretch of four or more consecutive U residues located between nt +6 and +25 relative to the cleavage site (Table 1). This frequency (26.7% of 180 polyadenylation sites) is much greater than would be expected for the occurrence of a 4-nt sequence on a random basis (8.9%, or 16 of 180 sequences that are 23 nt in length). It is likely that all of these sites can function as downstream elements that mediate binding of the 64K polypeptide, because they are all located within the permissible range established by the analysis of SVL derivatives. Many more of the downstream domains in our database contain degenerate versions of the U stretch as is the case for the Ad2L3 domain (Fig. 5A), and work is in progress to identify the variety of U-rich motifs that are able to sponsor binding of the 64K polypeptide (data not shown).

McLauchlan et al. (15) surveyed a series of viral, avian, and mammalian mRNA 3' termini and concluded that a downstream element with the consensus YGTGTTY is often involved in polyadenylation. This specific sequence is not present in any of the 180 mammalian sites that we identified, although many contained related GU-rich elements (data not shown). We suspect that GU-rich elements which are frequently found downstream of AAUAAA motifs can substitute both functionally and positionally for U-rich elements, but we have not tested this possibility. A relative lack of specificity of the 64K protein for a particular sequence motif might allow it flexibility in interacting with either U-rich or GU-rich downstream elements in the proper location. Alternatively, GU-rich elements may function by binding another, not yet identified factor.

We propose a model in Fig. 5B for the role of CstF in the accurate cleavage of pre-mRNA substrates during polyadenylation. This model suggests that a constituent of CPSF initially binds to the AAUAAA motif. Then CstF, which can interact with CPSF, binds through its 64K subunit to an appropriately positioned U-rich sequence downstream of the cleavage site on the RNA substrate. Thus, CPSF and CstF both interact with the pre-mRNA, and they interact with each other, spanning the polyadenylation cleavage site. CstF binding is essential for efficient pre-mRNA cleavage (7, 8, 27, 28). Efficient cleavage should require the presence of a 64K CstF protein binding site within an acceptable distance from the AAUAAA motif, and this prediction was borne out (Fig. 4B). Finally, movement of

the 64K protein binding site within a limited range might be expected to influence the location of the cleavage site. This prediction of the model also proved correct (Fig. 4D). It is interesting that the 64K polypeptide contains 12 repeated copies of a five-amino-acid unit (MEARA/G) that is predicted to form a stable  $\alpha$  helix (25). This rodlike structure, about 9 nm in length, might serve as a "molecular ruler," reaching from a contact with a component of CPSF to the U-rich downstream element on the pre-mRNA.

We conclude that polyadenylation factors contact pre-mRNA sequences both upstream and downstream of the cleavage site, and these sequences play a role in correct positioning of the polyadenylation machinery as well as choice of cleavage site.

#### ACKNOWLEDGMENTS

We thank J. Welch for help with the computer search, S. Berget, and G. Gilmartin for enlightening discussions, and Y. Takagaki and J. Manley for communication of results prior to publication.

This work was supported by grant CA38965 from the National Cancer Institute to T.S. and grant AI31165 to J.W. C.C.M. was a postdoctoral fellow of the American Cancer Society and an Associate of the Howard Hughes Medical Institute during portions of this work. T.S. is an American Cancer Society Professor and an Investigator of the Howard Hughes Medical Institute. J.W. is a Pew Scholar in the Biomedical Sciences.

#### REFERENCES

- Ahmed, Y. F., G. M. Gilmartin, S. M. Hanly, J. R. Nevins, and W. C. Greene. 1991. The HTLV-I Rex response element mediates a novel form of mRNA polyadenylation. *Cell* **64**:727-737.
- Chou, Z.-F., S. B. Amrute, and J. Wilusz. Unpublished data.
- Chou, Z.-F., F. Chen, and J. Wilusz. 1994. Sequence and position requirements for uridylylate-rich downstream elements of polyadenylation signals. *Nucleic Acids Res.* **22**:2525-2531.
- Conway, L., and M. Wickens. 1985. A sequence downstream of AAUAAA is required for formation of simian virus 40 late mRNA 3' termini in frog oocytes. *Proc. Natl. Acad. Sci. USA* **82**:3949-3953.
- Fitzgerald, M., and T. Shenk. 1981. The sequence 5'-AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. *Cell* **24**:251-260.
- Gil, A., and N. J. Proudfoot. 1987. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit  $\beta$ -globin mRNA 3'-end formation. *Cell* **49**:399-406.
- Gilmartin, G. M., and J. R. Nevins. 1989. An ordered pathway of assembly of components required for polyadenylation site recognition and processing. *Genes Dev.* **3**:2180-2189.
- Gilmartin, G. M., and J. R. Nevins. 1991. Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA. *Mol. Cell. Biol.* **11**:2432-2438.
- Keller, W., S. Bienroth, K. M. Lang, and G. Christofori. 1991. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal. *EMBO J.* **10**:4241-4249.
- MacDonald, C. C., and D. L. Williams. 1993. RNase H/oligodeoxyribonucleotide-directed mRNA purification (ROMP) of apoII mRNA. *Nucleic Acids Res.* **21**:765-766.
- Manley, J. L. 1988. Polyadenylation of mRNA precursors. *Biochim. Biophys. Acta* **950**:1-12.
- Mason, P. J., J. A. Elkington, M. M. Lloyd, M. B. Jones, and J. G. Williams. 1986. Mutations downstream of the polyadenylation site of a *Xenopus*  $\beta$ -globin mRNA affect the position but not the efficiency of 3' processing. *Cell* **46**:263-270.
- McDevitt, M. A., R. P. Hart, W. W. Wong, and J. R. Nevins. 1986. Sequences capable of restoring poly(A) site function define two distinct downstream elements. *EMBO J.* **5**:2907-2913.
- McDevitt, M. A., M. J. Imperiale, H. Ali, and J. R. Nevins. 1984. Requirements of a downstream sequence for generation of a poly(A) addition site. *Cell* **37**:993-999.
- McLauchlan, J., D. Gaffney, J. L. Whitton, and J. B. Clements.

1985. The consensus sequence YGTGTTY located downstream from the AAUAAA signal is required for efficient formation of mRNA 3' termini. *Nucleic Acids Res.* **13**:1347-1368.
16. **Moore, C. L., J. Chen, and J. Whorisky.** 1988. Two proteins crosslinked to RNA containing the adenovirus L3 poly A site require the AAUAAA sequence for binding. *EMBO J.* **7**:3159-3169.
17. **Murthy, K. G. K., and J. L. Manley.** 1992. Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *J. Biol. Chem.* **267**:14804-14811.
18. **Niwa, M., C. C. MacDonald, and S. M. Berget.** 1992. Are exons scanned during splice-site selection? *Nature (London)* **360**:277-280.
19. **Proudfoot, N. J., and G. G. Brownlee.** 1976. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature (London)* **263**:211-214.
20. **Ryner, L. C., and J. L. Manley.** 1987. Requirements for accurate and efficient mRNA 3' cleavage and polyadenylation of a simian virus 40 early pre-mRNA in vitro. *Mol. Cell. Biol.* **7**:495-503.
21. **Ryner, L. C., Y. Takagaki, and J. L. Manley.** 1989. Sequences downstream of AAUAAA signals affect pre-mRNA cleavage and polyadenylation in vitro both directly and indirectly. *Mol. Cell. Biol.* **9**:1759-1771.
22. **Sachs, A., and E. Wahle.** 1993. Poly(A) tail metabolism and function in eukaryotes. *J. Biol. Chem.* **268**:22955-22958.
23. **Sadofsky, M., and J. C. Alwine.** 1984. Sequences on the 3' side of hexanucleotide AAUAAA affect efficiency of cleavage at the polyadenylation site. *Mol. Cell. Biol.* **4**:1460-1468.
24. **Sadofsky, M., S. Connelly, J. L. Manley, and J. C. Alwine.** 1985. Identification of a sequence element on the 3' sides of AAUAAA which is necessary for simian virus 40 late mRNA 3' end processing. *Mol. Cell. Biol.* **5**:2713-2719.
25. **Takagaki, Y., C. C. MacDonald, T. Shenk, and J. L. Manley.** 1992. The human 64-kDa polyadenylation factor contains a ribonucleoprotein-type RNA binding domain and unusual auxiliary motifs. *Proc. Natl. Acad. Sci. USA* **89**:1403-1407.
26. **Takagaki, Y., and J. L. Manley.** 1994. Personal communication.
27. **Takagaki, Y., J. L. Manley, C. C. MacDonald, J. Wilusz, and T. Shenk.** 1990. A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes Dev.* **4**:2112-2120.
28. **Takagaki, Y., L. C. Ryner, and J. L. Manley.** 1989. Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes Dev.* **3**:1711-1724.
29. **Wahle, E., and W. Keller.** 1992. The biochemistry of 3'-end cleavage and polyadenylation of messenger RNA precursors. *Annu. Rev. Biochem.* **61**:419-440.
30. **Weiss, E. A., G. M. Gilmartin, and J. R. Nevins.** 1991. Poly(A) site efficiency reflects the stability of complex formation involving the downstream element. *EMBO J.* **10**:215-219.
31. **Wilusz, J., and T. Shenk.** 1988. A 64 kDa nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. *Cell* **52**:221-228.
32. **Wilusz, J., and T. Shenk.** 1990. A uridylate tract mediates efficient heterogeneous nuclear ribonucleoprotein C protein-RNA cross-linking and functionally substitutes for the downstream element of the polyadenylation signal. *Mol. Cell. Biol.* **10**:6397-6407.
33. **Wilusz, J., T. Shenk, Y. Takagaki, and J. L. Manley.** 1990. A multicomponent complex is required for the AAUAAA-dependent cross-linking of a 64-kilodalton protein to polyadenylation substrates. *Mol. Cell. Biol.* **10**:1244-1248.
34. **Zarkower, D., and M. Wickens.** 1988. A functionally redundant downstream sequence in SV40 late pre-mRNA is required for mRNA 3'-end formation and for assembly of a precleavage complex in vitro. *J. Biol. Chem.* **263**:5780-5788.