MOLECULAR AND CELLULAR BIOLOGY, July 1997, p. 3907–3914 Vol. 17, No. 7
0270-7306/97/$04.00+0
Copyright © 1997, American Society for Microbiology

RNA Recognition by the Human Polyadenylation Factor CstF

YOSHIO TAKAGAKI AND JAMES L. MANLEY*

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

Received 13 March 1997/Returned for modification 16 April 1997/Accepted 24 April 1997

Polyadenylation of mammalian mRNA precursors requires at least two signal sequences in the RNA: the nearly invariant AAUAAA, situated 5′ to the site of polyadenylation, and a much more variable GU- or U-rich downstream element. At least some downstream sequences are recognized by the heterotrimeric polyadenylation factor CstF, although how, and indeed if, all variations of this diffuse element are bound by a single factor is unknown. Here we show that the RNP-type RNA binding domain of the 64-kDa subunit of CstF (CstF-64) (64K RBD) is sufficient to define a functional downstream element. Selection-amplification (SELEX) experiments employing a glutathione *S*-transferase (GST)–64K RBD fusion protein selected GU-rich sequences that defined consensus recognition motifs closely matching those present in natural poly(A) sites. Selected sequences were bound specifically, and with surprisingly high affinities, by intact CstF and were functional in reconstituted, CstF-dependent cleavage assays. Our results also indicate that GU- and U-rich sequences are variants of a single CstF recognition motif. For comparison, SELEX was performed with a GST fusion containing the RBD from the apparent yeast homolog of CstF-64, RNA15. Strikingly, although the two RBDs are almost 50% identical and yeast poly(A) signals are at least as degenerate as the mammalian downstream element, a nearly invariant 12-base U-rich sequence distinct from the CstF-64 consensus was identified. We discuss these results in terms of the function and evolution of mRNA 3′-end signals.

Polyadenylation of mRNA precursors is a nearly ubiquitous posttranscriptional modification that occurs in the nuclei of all eukaryotic cells. The reaction involves a precise endonucleolytic cleavage of the pre-mRNA followed by synthesis of the poly(A) tail. This seemingly simple reaction requires an unexpectedly complex protein machinery consisting of a dozen or more distinct polypeptides (see references 14 and 19 for recent reviews). The first step in the reaction is specification of the poly(A) site. This has been studied in considerable detail in mammalian systems, where at least two signal sequences in the pre-mRNA, the nearly invariant AAUAAA and a much less conserved downstream GU (or U)-rich sequence, are required for efficient polyadenylation (reviewed in reference 27). Each of these elements is recognized by a distinct multisubunit protein, AAUAAA by cleavage-polyadenylation specificity factor (CPSF) and the downstream element by cleavage stimulation factor (CstF). Two less well-defined cleavage factors and poly(A) polymerase (PAP) appear to complete the complex. Some pre-mRNAs also contain sequence elements situated upstream of AAUAAA, which appear to work by diverse mechanisms. For example, the simian virus 40 (SV40) late pre-mRNA contains a sequence that can be recognized by the U1 snRNP A protein, which stabilizes CPSF binding to AAU AAA (16, 17), while in human immunodeficiency virus RNA, an upstream element seems to stabilize CPSF binding by making a second direct contact with CPSF (9).

CPSF and CstF are both multisubunit proteins, and each has been the subject of considerable study. CPSF was first identified as the only factor essential for both steps of the in vitro polyadenylation reaction [i.e., RNA cleavage and poly(A) synthesis (6, 7, 34)]. Consistent with the AAUAAA requirement for both steps, purified CPSF recognizes this element, albeit weakly (7, 15). The protein likely consists of four subunits, of 160, 100, 73, and 30 kDa (2, 11), although fully active preparations lacking the 30-kDa subunit have been isolated (9, 24). Early UV cross-linking studies suggested that the 160-kDa subunit (CPSF-160) directly contacts AAUAAA (1, 15), and in fact recombinant CPSF-160 can recognize AAUAAA-containing RNAs with specificity (25). cDNAs encoding the largest three subunits have been isolated, and none contain recognizable domains or similarities with other proteins (11–13, 25). CPSF-73 and -100, however, have significant similarity with each other, indicating they may have arisen from a common ancestor (13).

CstF is a heterotrimeric protein that consists of 77-, 64-, and 50-kDa subunits (35). The 64-kDa CstF subunit (CstF-64) was in fact the first polypeptide implicated in sequence-specific polyadenylation, based on UV cross-linking assays employing crude nuclear extracts (43). It was initially believed that CstF-64 might contact AAUAAA, as point mutations in the signal eliminated cross-linking. It is now clear that this reflects strong cooperative binding of CPSF and CstF, with the CPSF-RNA interaction being directly affected by mutations in AAU AAA (8, 24, 44). The strongest evidence that CstF contacts the downstream region comes from similar UV cross-linking experiments, which mapped the site of CstF-64 cross-linking to short U-rich sequences within the downstream regions of two different pre-mRNAs (18). cDNAs encoding all three subunits have been cloned and display interesting features. CstF-50 contains seven transducin, or WD-40, repeats, suggestive of protein-protein interactions (37). Although essential for CstF function in vitro, its role is unknown. CstF-77 is the apparent homolog of the *Drosophila* suppressor-of-forked protein and functions to bridge the other two CstF subunits (38). It also interacts strongly with CPSF-160, which is likely responsible at least in part for the cooperative RNA binding displayed by the two factors (25). CstF-64 contains an N-terminal RNP-type RNA binding domain (RBD) and an unusual C-terminal region consisting of a 300-residue 40% P+G region that is interrupted by 12 tandem copies of the pentapeptide consensus MEARA/G (36). The functions of these domains are unknown, although it seems likely that the RBD is involved in binding the pre-mRNA. The intracellular levels of CstF-64 have recently been shown to be important in the switch from membrane-bound to secreted immunoglobulin M H-chain mRNA that occurs during B-cell differentiation (39). This reg-

* Corresponding author.

ulation appears to reflect the differential affinities of CstF for the GU-rich regions of the two poly(A) sites.

Additional insights into how these factors might function to specify poly(A) sites is beginning to emerge from studies with *Saccharomyces cerevisiae*. Recent work has revealed surprising similarities between yeast polyadenylation factors and components of mammalian CPSF and CstF (reviewed in reference 20). This was unexpected because poly(A) signals in yeast pre-mRNAs do not bear a clear resemblance to those in mammals, and it is even difficult to define convincing consensus sequences (reviewed in reference 10). The existence of yeast homologs of components of both CstF (23, 38) and CPSF (3, 13, 32) suggests, however, that important aspects of the RNA-protein and protein-protein interactions important for recognizing and forming mRNA 3′ ends are conserved throughout eukaryotes. However, little is yet understood about these interactions in yeast. For example, it is not yet known what RNA elements are recognized by which proteins.

In this paper, we investigate some of the RNA-protein interactions involved in polyadenylation using the SELEX methodology. We show that the CstF-64K RBD is sufficient to recognize consensus motifs that closely resemble both the GU-rich and U-rich sequences found in vivo. These sequences bind intact CstF with high affinity and function as downstream elements in CstF-dependent in vitro cleavage assays. Strikingly, given the variable nature of yeast RNA signals, the RBD of the apparent yeast CstF-64 homolog, RNA15, selects essentially a single U-rich sequence. This consensus bears only limited resemblance to the 64K RBD-selected sequences and does not interact with CstF, but it does reveal some similarity to a recently suggested consensus element in yeast pre-mRNAs. We discuss the significance of these results with respect to poly(A) site specification.

MATERIALS AND METHODS

Preparation of GST-RBD fusion proteins. To prepare a plasmid encoding the glutathione *S*-transferase (GST)–64K RBD fusion protein, an *Nde*I site was first created at the translation initiation codon in the human CstF-64 subunit cDNA (pZ64-18 [36]) by oligonucleotide-directed mutagenesis using synthetic oligonucleotide 64K-NdeI (5′-CTCAACAGACATATGGCGGG-3′). An *Nde*I-*Ban*I fragment derived from this plasmid, which encompasses the N-terminal 108 amino acids, including the RBD (amino acid residues 17 to 96), was treated with the Klenow fragment of DNA polymerase I and inserted into the *Sma*I site of the pGEX-2T vector (31) to prepare pGEX-64K RBD. To prepare a plasmid encoding the GST-RNA15 RBD fusion protein, a DNA fragment encompassing the N-terminal 109 amino acids of yeast RNA15, which includes the RBD (amino acids 19 to 96) (22) was obtained by PCR using synthetic oligonucleotides RNA15-5′ (5′-TATGAATAGGCAGAGCGGTG-3′) and RNA15-3′ (5′-TACT GTTGTTGTTGCTGTTGTG-3′) as primers and 50 ng of yeast genomic DNA as a template. Primers and template were denatured at 94°C for 1 min and annealed at 60°C for 1 min, and DNA was synthesized at 72°C for 1 min for 35 cycles. The 330-bp DNA was purified on a 1% agarose gel, treated with T4 polynucleotide kinase and Klenow fragment, and inserted into the *Sma*I site of the pGEX-2T vector to prepare pGEX-RNA15 RBD. Sequences of the encoded fusion proteins were verified by DNA sequencing.

To express the GST-RBD fusion proteins, *Escherichia coli* JM101 cells transformed with pGEX-2T, pGEX-64K RBD, or pGEX-RNA15 RBD were grown in 250 ml of 2 × YT containing 200 mg of ampicillin per liter to an optical density at 600 nm of 0.5 at 37°C, and protein expression was induced with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at room temperature for 2 h. After centrifugation at 6,000 rpm and 4°C for 10 min in a Sorvall GS-4 rotor, bacterial pellets were resuspended in 20 ml of ice-cold sonication buffer containing 50 mM Tris-HCl (pH 7.9), 100 mM NaCl, 1 mM EDTA, 1 mM dithiothreitol (DTT), 1 mM phenylmethylsulfonyl fluoride (PMSF), 0.5% (vol/vol) Nonidet P-40, and 10% (vol/vol) glycerol and sonicated with a sonifier (Branson) at setting 2 three times for 30 s each on ice. Sonicated bacterial suspension was centrifuged at 15,000 rpm and 4°C for 20 min in a Sorvall SS34 rotor, and the supernatant was loaded on a 1-ml column of glutathione (GSH)-agarose (Sigma) equilibrated with sonication buffer in the cold room. After washing with sonication buffer and wash buffer (same as sonication buffer but no Nonidet P-40), GST-RBD fusion proteins were eluted with 20 mM GSH in wash buffer and dialyzed twice against dialysis buffer containing 20 mM HEPES-NaOH (pH 7.9), 100 mM NaCl, 0.2 mM EDTA, 0.5 mM DTT, 0.5 mM PMSF, and 20% (vol/vol) glycerol. Protein concentrations were determined on a sodium dodecyl sulfate–10% polyacrylamide gel by staining with Coomassie brilliant blue using bovine serum albumin as a standard.

Selection-amplification (SELEX) of RNAs with high affinities for RBDs. SELEX was carried out as described previously (33, 40) with modifications. First, template DNA (0.13 μg) containing 20 nucleotides (nt) of randomized sequence in the middle was amplified by PCR using T7 (1.5 μg) and Rev (1.3 μg) primers. The first five cycles of PCR were carried out by denaturing at 94°C, annealing at 50°C, and synthesizing DNA at 72°C for 30 s each. The next 30 cycles were done under the same conditions, except annealing was at 65°C. One microgram of the PCR products was digested with *Bam*HI and transcribed with 40 U of T7 RNA polymerase in the presence of [α-^{32}P]UTP as a tracer. RNA transcripts were purified on an 8.3 M urea–5% polyacrylamide gel, eluted, and recovered by ethanol precipitation in the presence of 10 μg of tRNA. For the enrichment of RNAs with high affinities for the GST-RBD fusion proteins, 10 μg of the fusion proteins was conjugated with 20 μl of GSH-agarose in 200 μl of sonication buffer by gently mixing on a Nutator (Clay Adams) for 1 h in the cold room. After being washed three times with sonication buffer and binding buffer containing 8 mM HEPES-NaOH (pH 7.9), 40 mM NaCl, 2 mM EDTA, 0.2 mM DTT, 0.2 mM PMSF, and 8% (vol/vol) glycerol for 5 min each in the cold room, GST-RBD fusion protein–GSH-agarose conjugates were mixed with 80 μl of binding buffer containing 3.2 μg of *E. coli* RNA and 1 μg of the in vitro-transcribed RNAs. After incubation at 30°C for 10 min with occasional shaking, the conjugates were washed five times for 5 min each with binding buffer at room temperature; resuspended in 200 μl of proteinase K buffer containing 20 mM Tris-HCl (pH 7.9), 100 mM NaCl, 10 mM EDTA, 1% (wt/vol) sodium dodecyl sulfate, 15 μg of tRNA, and 70 μg of proteinase K; and incubated at 30°C for 30 min with occasional mixing. The mixtures were extracted with 200 μl of phenol-chloroform, phenol-chloroform was back-extracted with 50 μl of Tris-EDTA (pH 7.9), and RNAs were recovered from the combined aqueous phase by precipitation with 2 volumes of ethanol in the presence of 0.3 M NaCl. RNA pellets were washed with ethanol, dried, and resuspended in H_2O. RNAs were reverse transcribed with avian myeloblastosis virus reverse transcriptase (Life Sciences) with 0.64 μg of Rev primer in a total volume of 50 μl at 42°C for 1 h. Reaction mixtures were extracted with 50 μl of phenol-chloroform, and cDNAs were recovered by ethanol precipitation in the presence of 2 M ammonium acetate and amplified by PCR as described above. SELEX was repeated six to eight more times. A fraction of PCR products from rounds 7 to 9 was digested with *Bam*HI and *Eco*RI, subcloned into the pBluescript SK vector, and subjected to sequence analysis using Sequenase (U.S. Biochemicals) and the T3 primer.

Gel shift assays. To prepare RNAs for gel shift assays, pBluescript SK plasmids containing the selected sequences were digested at the *Acc*65I site in the vector and at the *Bam*HI site at the 3′ ends of the SELEX templates. Templates were transcribed with T7 RNA polymerase in the presence of [α-^{32}P]ATP and purified on denaturing polyacrylamide gels as above. For gel shift assays, in vitro-transcribed RNAs (~59 nt, 1.5 nM) were incubated with 0.18 to 7.5 μM GST-64K RBD fusion protein in 12.5 μl of binding buffer containing 0.5 μg of *E. coli* RNA at 30°C for 10 min. The reaction mixtures were loaded on 5% nondenaturing polyacrylamide gels (acrylamide-bis, 40:1) containing 0.5× Tris-borate-EDTA and 1% (vol/vol) glycerol and electrophoresed at 2 W for 1 h in the cold room. For gel shift assays with intact CstF, 1.5 nM RNAs were incubated with 1, 3.3, or 10 nM CstF purified from HeLa cell nuclear extracts (35) (purity, >90%) in a buffer containing 20 mM $(NH_4)_2SO_4$ instead of NaCl. After incubation with 5 mg of heparin per ml for 10 min on ice, the reaction mixtures were electrophoresed on 4% nondenaturing gels (acrylamide-bis, 80:1) containing 0.5× Tris-borate-EDTA in the cold room. Apparent K_ds were determined by estimating the protein concentration required to shift 50% of the RNA probe. This was monitored by the disappearance of unbound probe, rather than accumulation of RNA-protein complexes, because the latter appear to dissociate during gel electrophoresis.

In vitro cleavage reactions. To prepare RNA substrates for in vitro cleavage reactions, the downstream sequence of the adenovirus type 2 (Ad2) L3 poly(A) site (28) was replaced with selected sequences by inserting *Eco*RI (filled-in)-*Bam*HI fragments derived from SELEX templates between the *Ava*I [filled-in, 9 bp downstream of the poly(A) site] and *Bam*HI (in the vector) sites in pG3L3-A. After *Bam*HI digestion, linearized plasmid DNAs were transcribed with SP6 RNA polymerase, and RNA transcripts were purified by denaturing polyacrylamide gel electrophoresis as above. Capped RNA substrates (1.5 nM) were incubated in reaction mixtures (12.5 μl) containing 1 μl of HeLa cell CPSF (Mono Q), 1 μl of CstF (Mono S; 0.8 or 4.0 nM), 1 μl of PAP (Mono S), and 2 μl of cleavage factor I (CFI) plus CFII (Mono Q) at 30°C for 1.5 h. Reaction products were purified and fractionated on 8.3 M urea–5% polyacrylamide gels (see references 29, 34, and 35 for details of protein purification and reaction conditions).

RESULTS

The CstF-64 RBD selects GU-rich sequences. To investigate the specific RNA-protein interactions involved in poly(A) site recognition, we designed experiments to examine the RNA binding specificity of polypeptides involved in this process. As mentioned in the introduction, previous studies have provided

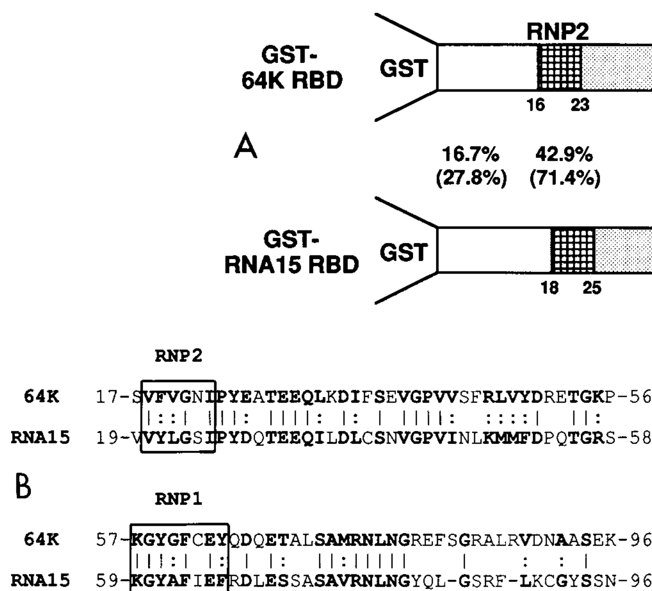

FIG. 1. Structures of GST-RBD fusion proteins. (A) Primary structures of GST-64K RBD (top) and GST-RNA15 RBD (bottom) fusion proteins are compared, and amino acid identities and similarities (in parentheses) of each region are shown. RNP1 and RNP2 consensus sequences and other parts of the RBDs are indicated by hatched boxes and dotted boxes, respectively, and the last residue of each domain is numbered. (B) The amino acid sequence of 64K RBD (top) is optimally aligned with that of RNA15 RBD (bottom) by using the FASTA program (26), and amino acids of each protein are numbered on the sides. Identical and similar residues between these two RBDs are indicated by lines and dots, respectively, and are shown in boldface. RNP1 and RNP2 consensus sequences are boxed. Note that spacings between the RNP2 and RNP1 consensus sequences are exactly the same between 64K RBD and RNA15, while the C-terminal domain of RNA15 RBD is shorter than that of 64K RBD by 2 residues.

support for the idea that a downstream sequence element is recognized by CstF-64. To investigate this further, we decided to perform SELEX experiments (41) with a derivative of CstF-64. Specifically, it seemed logical that RNA binding might be mediated by the RBD. Indeed, previous experiments showed that a 250-residue N-terminal fragment encompassing the RBD could bind RNA, although sequence specificity was not demonstrated (36). We therefore constructed a GST fusion protein containing residues 1 to 108 of CstF-64 (Fig. 1) (the RBD extends approximately from residues 17 to 96). The fusion protein, GST-64K RBD, was expressed in and purified from *E. coli*, bound to GSH-agarose beads, and used in SELEX experiments as described in Materials and Methods. The RNA population used consists of molecules 59 nt long, with 20 nt of randomized sequence (33).

SELEX was performed for seven cycles, after which the sequences of 23 cloned fragments were determined (Fig. 2). Strikingly, all of the sequences were characterized by the presence of long GU-rich motifs (boldfaced in Fig. 2). Several features of the selected sequences are particularly noteworthy and immediately suggest similarities with authentic downstream elements. First, although all sequences are characterized by a high GU content, there is no consensus sequence, which is also the case with downstream elements. Second, the selected GU motifs are characterized by GU dinucleotides, frequently repeated, and/or by U repeats (up to U_4), but there are no G repeats. This is consistent with previous studies that mapped the site of UV-induced cross-linking of CstF-64 (in nuclear extracts) to U-rich regions in two pre-mRNAs (18). These authors also pointed out that a significant fraction of downstream sequences contain a U-rich motif, analogous to what is observed in our selected sequences. Furthermore, these results are consistent with our preliminary experiments (not shown) indicating that GST-64K RBD could bind poly(U) but not homopolymers of any of the other three nucleotides. It is noteworthy that most of the selected sequences contain C residues 5′ to the GU-rich sequences. However, the number of C residues found (three to seven) and the spacing between the C residues and the GU-rich sequences (zero to four) are both variable. Furthermore, the random sequences, used as negative controls (R5 and R10), were also C rich (7 and 9 of 19 residues, respectively) but did not bind GST-64K RBD in RNA gel shift assays (see below), suggesting that the C-rich sequences do not constitute part of the consensus.

We next verified that the selected sequences are capable of

Random sequences

R 5	CUCUUGAGCAUCGCUGUCC
R10*	CUCCACGCUGAACUCUGCC

Selected sequences

#12	GCCCCACUGUCUUA**UUGUGU**
#11*	GGCCCGA**GUGUGUGU**AUGC
#14	GGCCCUC**GUGUGUGU**AUUGC
#10	CAUGCCCCUCUGUA**GUUG**C
#17*	AGCCCCC**UGUUGU**CUGC
#18	UUGCCCCCCC**UGUUGU**CUGC
#20	UUGGCUCC**UGUUGU**CUUCGC
#15*	AGUGCCCC**UGUGUUUG**C
# 3	UAUGCCCCCC**UGUGUGUUGU**
# 4	GUCCCCA**UUGUGUGUUGU**
#21*	UGGUCCC**UUGUGUGUUG**C
# 2*	GGUCCCCAA**GUGUGUUUUG**C
#22*	UGGCCC**UUGUGUGUGUUG**C
# 7	UAGGCCC**GUGUGUGUUUGU**
#19*	UGGCCCA**GUGUGUGUUUG**C
#23	UGGUCCC**GUGUGUGUUUG**C
#13	UGCCCCCC**UGUUGUUUG**C
#16*	UGGCCUCC**GUUGUGUUUG**C
# 5*	AGCCCC**GUUUGUGUUUG**C
#24	AGGUCCCA**GUUGUGUGUUGU**
# 8	GGCUCC**GUGUGUUAGUGU**
# 1	GGCUCC**GUGUGUUAGUGUGU**
# 6	AGCCCC**UGUGUUCGUUUUG**C

FIG. 2. 64K RBD selects GU-rich sequences from a pool of RNAs with randomized sequences. Sequences of RNAs selected by SELEX experiments using GST-64K RBD fusion protein as a ligand (see Materials and Methods) are shown in order of increasing complexity from the top downwards and are numbered on the left. GU-rich sequences found in all of the selected RNAs are shown in boldface. RNAs used for gel shift assays (Fig. 3 and 4) are indicated by asterisks.

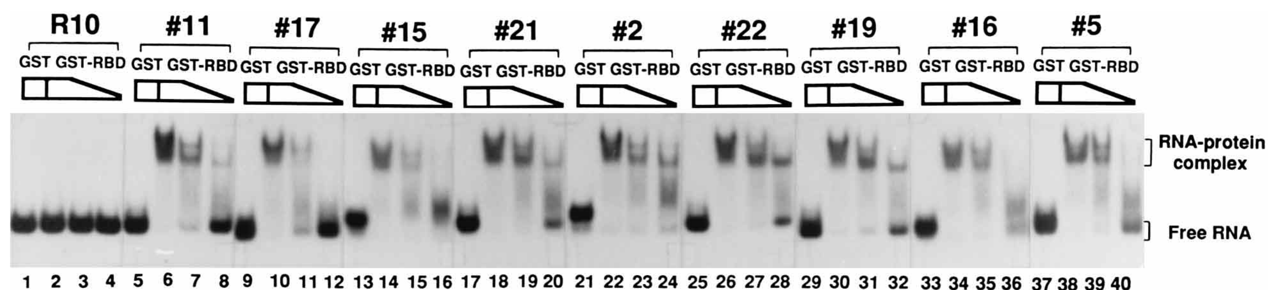

FIG. 3. GST-64K RBD specifically binds RNAs with GU-rich sequences. After incubation of RNAs containing a random sequence (lanes 1 to 4) or selected GU-rich sequences (the rest) with 7.5 μM GST or decreasing amounts of GST-64K RBD fusion protein (GST-RBD; 7.5, 1.5, or 0.3 μM), free RNAs and RNA-protein complexes were separated on a nondenaturing polyacrylamide gel. RNAs used are numbered on top, and positions of free RNAs and RNA-protein complexes are indicated on the right.

binding GST-64K RBD in a different assay, i.e., RNA gel shift assays. An example of the data is presented in Fig. 3. Random RNAs transcribed from clones derived from the starting pool of PCR products (two were tested) failed to bind GST-64K RBD detectably at any concentration tested (e.g., lanes 1 to 4). In contrast, all 20 of the selected sequences tested bound the protein (e.g., lanes 5 to 40). There were, though, differences in apparent affinities among the selected sequences. For example, sequence 2 (lanes 21 to 24) showed the strongest binding, with essentially all of the RNA shifted at the lowest concentration of GST-64K RBD tested (180 nM), while at most about 10% of sequence 17 was bound at this concentration. Additional gel shift experiments with lower concentrations of GST-64K RBD (not shown) indicate a range of apparent K_ds from ~100 nM (sequence 2) to ~1.5 μM (sequence 17). (See Materials and Methods for the method of K_d calculations.) Many of the samples showed significant smearing, especially at the lowest concentrations tested. This likely reflects dissociation of complexes during electrophoresis. Careful examination of Fig. 3 suggests that two types of RNA-protein complexes with different mobilities were formed. Since CstF-64 does not appear to form dimers (not shown), these seem to reflect formation of GST-64K RBD dimers mediated by GST dimerization (30a) and/or binding of more than one GST-64K RBD molecule to a single RNA. We do not know the basis for the differential affinities of the RNAs, although it is noteworthy that the highest-affinity binder, sequence 2, is the only sequence tested containing a U_4 motif.

Purified CstF binds specifically to selected sequences. The above-described experiments demonstrate that the CstF-64 RBD is sufficient to select RNA sequences that resemble natural downstream elements. But are these sequences bound selectively by intact CstF? To address this question, we performed gel shift assays similar to those in Fig. 3, except that CstF purified from HeLa cells (35) was used. Figure 4 displays the results obtained with the same RNAs as in Fig. 3, and the results were qualitatively extremely similar. Two random RNAs bound barely detectable levels of CstF (e.g., lanes 1 to 4) and did so only at the highest protein concentrations tested (~10 nM), while all of the selected RNAs displayed significant binding. With several RNAs, 100% of the probe was shifted at 10 nM CstF. Differences were again detected in the apparent

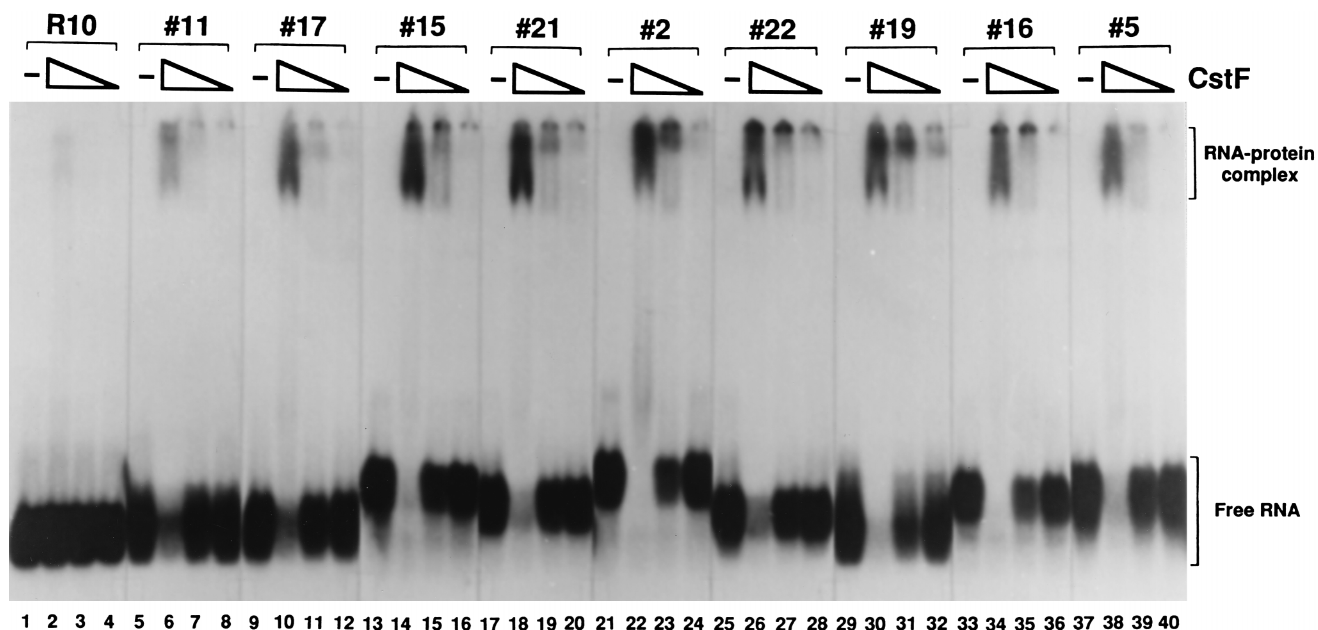

FIG. 4. Purified CstF complex specifically binds RNAs with GU-rich sequences. After incubation of RNAs with decreasing amounts of purified CstF (purity >90%; 10, 3.3 or 1.0 nM) or without CstF (−), free RNAs and RNA-protein complexes were separated as in Fig. 3.

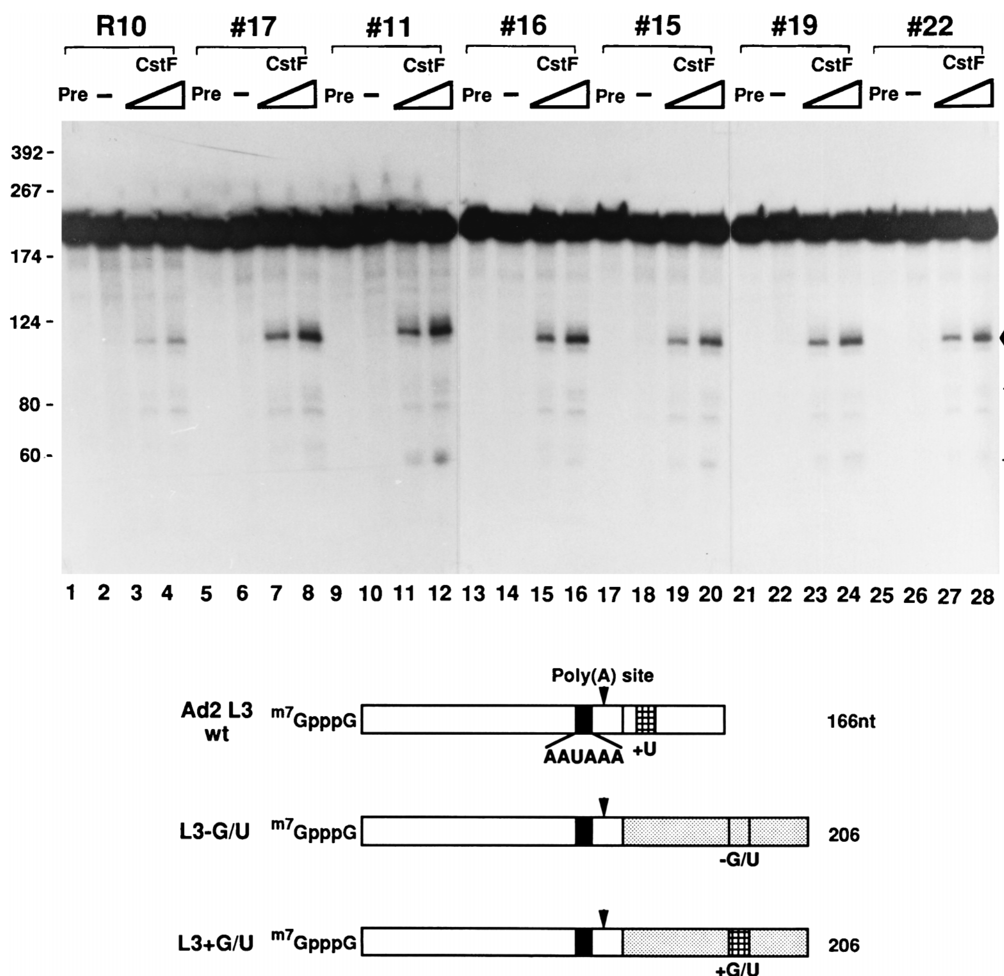

FIG. 5. GU-rich sequences function as downstream elements in in vitro cleavage reactions. Chimeric pre-mRNAs (see bottom) containing a random sequence (lanes 1 to 4) or selected GU-rich sequences (the rest) situated downstream of the Ad2 L3 poly(A) site were incubated with partially purified CPSF, CFI, CFII, and PAP in the absence (−) or presence of increasing amounts (0.8 or 4.0 nM) of purified CstF, and the reaction products were fractionated on a denaturing polyacrylamide gel. Positions of DNA size markers (in nucleotides) are indicated on the left, and those of the upstream and downstream cleavage products are indicated by an arrow and a bracket, respectively, on the right. The downstream region of the Ad2 L3 gene (Ad2 L3 wt) was replaced with SELEX templates containing a random sequence (L3−G/U) or selected GU-rich sequences (L3+G/U) (see Materials and Methods). Sequences derived from the Ad2 L3 gene and SELEX templates are shown by open and dotted boxes, respectively, and GU-rich and U-rich sequences are indicated by hatched boxes. The lengths of pre-mRNAs are indicated on the right.

affinities of selected sequences, which paralleled those observed with GST-64K RBD: sequence 2 displayed the highest affinity (apparent $K_d \sim 1.5$ nM), and sequence 17 displayed the lowest (~5 nM). It is noteworthy that the differences in apparent K_d observed with the purified factor were significantly less than those with the recombinant fusion protein (~3-fold versus 10- to 20-fold), and more importantly, the apparent affinity of CstF for the selected sequences was at least 50-fold greater than that of CstF-64 RBD. We discuss below possible explanations for these differences. However, the most important conclusion from these experiments is that sequences selected by the RBD of CstF-64 are recognized specifically and with high affinity by intact CstF.

Selected sequences function as downstream elements in reconstituted CstF-dependent cleavage reactions. As a final test of the functional significance of the selected sequences, we tested their ability to function in 3′ cleavage assays. For this we used a pre-mRNA containing the well-studied adenovirus L3 poly(A) site (28, 30). Reactions were performed with the required purified or partially purified components, CPSF, CFI, CFII, and PAP, and increasing concentrations of CstF. As we and others have observed previously (28), significant cleavage can occur in the absence of a downstream element. This can be seen in the samples that contained the RNA with one of the random sequences inserted 3′ to the site of RNA cleavage (Fig. 5, lanes 1 to 4). It is noteworthy that this cleavage is entirely dependent on CstF, despite the lack of a binding site (see Discussion). However, insertion of any of six different selected sequences resulted in significant enhancement (three- to fivefold) of cleavage (lanes 5 to 28). Somewhat unexpectedly, the extent of enhancement was similar with all six sequences. Thus, sequences 17 and 11, which had the lowest apparent affinities for CstF, were at least as efficient as sequences 15 and 16, which had among the highest. (For technical reasons, we were unable to test sequence 2, which had the highest apparent affinity.) Sequence 16 is also noteworthy because it bears a particularly strong similarity to well-characterized natural sequences (21) (see Discussion). Similar results were obtained when the different substrates were processed for shorter times or different concentrations of factors were employed (results not shown). These results indicate that sequences selected by

7 rounds

x13	GAC**UGUGUAUUCUCC**GGAU
x 6	GUA**UGCGUAUUUUCC**UGGU
x 2	CAAUGCC**UGCGUAUUCUCC**G
x 1	GAC**UGUGUAUUCUCC**G

9 rounds

x 9	GUA**UGCGUAUUUUCC**UGGU
x 7	GAC**UGUGUAUUCUCC**GGAU
x 4	CAAUGCC**UGCGUAUUCUCC**G
x 1	GUA**UGCGUAUUCUCC**UGGU

Consensus **UGCGUAUUCUCC**
U U

FIG. 6. RNA15 RBD selects U-rich sequences from a pool of RNAs with randomized sequences. Sequences of RNAs selected by SELEX experiments using GST-RNA15 RBD fusion protein as a ligand are shown in order of decreasing frequency from the top downwards. Sequences obtained after seven and nine cycles are shown. The number of RNAs which contain each sequence is indicated on the left. U-rich sequences present in all of the selected RNAs are shown in boldface, and the consensus sequence is shown at the bottom.

the CstF-64 RBD can provide downstream sequence function in a CstF-dependent cleavage assay.

The yeast RNA15 RBD selects a single U-rich consensus. The apparent yeast homolog of CstF-64 is RNA15 (38). The RBDs of the two proteins are 43% identical and 63% similar (Fig. 1). Although RNA15 has been shown to participate in the yeast 3′-end formation reaction (23), almost nothing is known about its RNA binding properties. Furthermore, given the degeneracy of yeast polyadenylation signals, it is difficult to predict how RNA15 interacts specifically with the pre-mRNA. To begin to study RNA recognition by RNA15 and to compare the RNA binding specificity of two closely related proteins, we performed SELEX with a GST-RNA15 RBD fusion protein (Fig. 1). As the apparent affinity of the protein for the RNA pool was lower than with GST-64K RBD, we initially performed nine rounds of SELEX prior to sequencing. However, RNAs obtained after seven rounds were also cloned and sequenced, with results indistinguishable from those obtained after nine cycles (Fig. 6). The most striking feature of the selected sequences is that they are nearly identical and give rise to a clear consensus. Not only does this contrast with the degenerate nature of the sequences selected by GST-64K RBD, but it is all the more unexpected given the lack of a convincing consensus among yeast poly(A) sites. The RNA15 RBD consensus is U rich, which is also a feature of the 64K RBD-selected sequences. Interestingly, almost half of the RNA15 RBD-selected RNAs contain $UGUGU\underline{A}U_2$, which is nearly identical to a $UGUGU\underline{G}U_2$ sequence found in about one-third of the 64K RBD-selected RNAs (Fig. 2, no. 3, 4, 7, 19, and 21 to 24). However, although GST-RNA15 RBD binds an RNA containing its selected sequence in a gel shift assay, GST-64K RBD showed no affinity for this RNA (results not shown). It is noteworthy that the GST-RNA15 RBD consensus bears some similarity to an upstream AU-rich efficiency element suggested by Guo and Sherman (10). In any event, it is clear that these two closely related RBDs recognize somewhat similar but distinct sequences, in keeping with the significant differences between mammalian and yeast polyadenylation signals.

DISCUSSION

The data presented here establish that the RBD of CstF-64 is sufficient to define the sequence specificity of the intact CstF complex. Although the specificities of the GST-64K RBD fusion protein and native CstF were found to be similar, the apparent dissociation constants of the RNA-protein complexes found with the latter were as much as a factor of 100 lower than those found with the former. Our data also provide a parsimonious explanation for results from previous studies that suggested the existence of two distinct classes of downstream elements, GU rich and U rich. Below we discuss these conclusions, as well as comment on how our results bear on the relationship between poly(A) signals in yeast and mammals.

Previous studies suggested that CstF is responsible for recognition of at least some downstream elements and that CstF-64 directly contacts the pre-mRNA (18). But it has been less clear that the factor is by itself capable of high-affinity, sequence-specific binding. For example, gel shift experiments with purified CstF preparations and several natural pre-mRNAs revealed only weak interactions, but these could be stabilized by addition of CPSF (8, 24). Likewise, efficient UV cross-linking of CstF-64 to pre-mRNA was found to require CPSF in addition to CstF (8, 24, 44). Also, recombinant CstF-64 was shown to UV cross-link to an SV40 late pre-mRNA without detectable specificity (18). These results together had led to the view that both high-affinity and sequence-specific binding may require interactions with additional polyadenylation factors, e.g., CPSF. Our results are thus important because they establish that CstF-64 and specifically its RBD have the intrinsic ability to recognize functional, GU-rich RNA sequences. Equally important, CstF is capable of recognizing these sequences with high affinity. As the sequences that we have identified here match well the natural sequences present in pre-mRNAs studied in previous experiments (see below), it is not clear why high-affinity binding by CstF was not observed previously. One possibility is that the RNAs used in our experiments were three- to fourfold shorter than RNAs employed previously. Consistent with this, we recently examined CstF binding to immunoglobulin M H-chain secreted and membrane form poly(A) sites, using gel shift assays very similar to those described here, except that the RNA length was significantly greater (39). Although both RNAs contain GU-rich regions that resemble those identified here, 10- to 50-fold greater protein concentrations were required to obtain detectable shifted complexes. If RNA size is in fact a significant determinant in establishing complex stability, then in vivo it is likely that CPSF-CstF interactions are required for poly(A) site specification, as has been suggested previously. However, our results are important because they indicate that CstF, via the CstF-64 RBD, has the ability to bind RNA specifically and with high affinity.

Over a decade ago, McDevitt et al. proposed the existence of two distinct downstream elements, denoted GU rich and U rich (21). Although GU richness seems to be more characteristic of the majority of downstream elements, support for the importance of U-rich regions has come from more recent studies establishing that short U stretches can provide downstream function in in vitro cleavage assays (5, 18, 45). Based on apparently distinct properties of the two sequences, it was suggested that the elements might be recognized by distinct *trans*-acting factors. Our results, however, provide a strong argument that both sequences are bound by CstF-64. First, a general feature of the selected sequences is that, while being overall GU rich, they also frequently contain U_n (but not G_n) stretches, and thus many may be called both GU rich and U rich. Second, stronger evidence comes from a direct comparison of specific sequences. McDevitt et al. identified two naturally occurring octamers that were sufficient to reconstitute downstream function in an in vivo polyadenylation assay, a GU-rich sequence (5′-GUUGUGGU) from SV40 and a U-rich sequence (5′-UUGUUUUU) from adenovirus. Both se-

quences match many of our selected sequences (six of eight or better). However, comparison of one, sequence 16, which contains a GU-rich decamer and displays very high affinity for CstF, is particularly informative. Specifically, sequence 16 contains seven of eight matches to both natural sequences. McDevitt et al. also analyzed a number of mutations in each element. Importantly, mutation of either G in the GG dinucleotide to U enhanced activity threefold. This finding is in keeping with our failure to detect G repeats in our selected sequences. Furthermore, one of the GG mutations created a perfect match with sequence 16. Remarkably, a single U-to-G mutation in the adenovirus U-rich sequence, which was without effect on activity (21), also created a perfect match with sequence 16. These results on the one hand strongly support the physiological relevance of our selected sequences while on the other argue that the two classes of sequences initially identified by McDevitt et al. are both actually CstF-64 binding sites.

The U-rich sequences display another characteristic feature: they are inactive when situated more than about 30 bases downstream of the cleavage site (5, 18, 21). In contrast, GU-rich elements are active up to 50 to 60 bases from the cleavage site (in our assays, the GU-rich elements were nearly 55 bases downstream). We speculate that this difference reflects a lower affinity of CstF-64 for the U-rich sequences tested. This leads to two important, related questions: what determines the strength of a CstF-64 binding site, and how does this contribute to the efficiency with which the poly(A) site is used? Our results provide a framework with which to consider these issues but do not provide hard and fast answers. Our data suggest only that a GU-rich sequence lacking consecutive G's is important for high-affinity binding and are consistent with the length of the sequence being an important factor. But beyond this it is not yet possible to say what specific sequences, if any, constitute an optimal CstF recognition site. It is also possible that flanking sequences, perhaps via secondary or tertiary interactions, influence binding.

How does binding affinity ultimately influence cleavage efficiency? We were surprised to find no significant differences in our reconstituted processing assay between high- and low-affinity sequences. We suspect that this reflects the strong cooperative interaction between CPSF and CstF (8, 24), which likely compensates for at least relatively small differences in CstF affinity. In fact, our data show that detectable, CstF-dependent 3′ cleavage can occur in the absence of any recognizable downstream motif. We believe that this reflects an interaction of CstF-64 with sequences that have only a very slight resemblance to a high-affinity binding site and that this interaction is stabilized by cooperative binding with CPSF. Although we obtained similar results in the in vitro cleavage of RNAs with different downstream GU-rich sequences both in nuclear extracts (not shown) and in the reconstituted reaction system (Fig. 5), the strength of the CstF-RNA interaction may play a more important role in vivo, where CstF may have to compete with other RNA binding proteins not present in the in vitro system.

An unexpected aspect of our experiments is the striking difference in apparent affinities of GST-64K RBD and CstF for the selected sequences. We cannot rule out the possibility that this reflects an experimental artifact, for example, that a large fraction of the recombinant protein was inactive. However, we have no indication that this was the case, and the magnitude of the effect leads us to suspect that some property of CstF greatly enhances binding affinity without affecting specificity. An intriguing possibility is that the MEARA/G repeat structure in CstF-64 contacts the pre-mRNA, perhaps via ionic interactions between arginine residues and the phosphate backbone. Although we currently have no data addressing this possibility, we have been unable to detect a role for this region of CstF-64 in protein-protein interactions (unpublished data).

Our results with GST-RNA15 RBD indicate that its RNA binding specificity is distinct from that of GST-64K RBD. This is in keeping with the striking differences in *cis*-acting sequences required for 3′-end formation in yeast and mammals. Guo and Sherman (10) recently described possible consensus sequences for yeast mRNA 3′-end formation. Specifically, they proposed the existence of two key elements, an A-rich positioning element located frequently 10 to 30 bases upstream of the cleavage site and an AU-rich efficiency element situated a variable distance 5′ to the positioning element. Although considerable progress has been made recently in identifying *trans*-acting factors necessary for yeast polyadenylation (reviewed in references 20 and 42), little is known about the RNA-protein interactions involved. Our data support our previous suggestion (20) that RNA15, which is a component of CFI (4, 23), is involved in recognition of the upstream efficiency element, by virtue of the U richness of the selected sequence that we identified. Although there are no perfect matches to any of several suggested AU-rich hexanucleotide efficiency elements (10), our consensus sequence displays four or five of six matches to all of them. (By contrast, there were at best only two matches with any of the suggested positioning element sequences [10]). It is unclear, however, why our selected sequence forms such a nearly invariant consensus sequence. This is unlike the extremely variable nature of efficiency elements and also distinct from the variation observed with the GST-64K RBD selected sequences. Whatever the explanation for this is, our data have shown that two closely related RBDs can have quite distinct RNA binding preferences. This difference parallels the divergence in polyadenylation signal sequences between yeast and mammals. An important conclusion that emerges from our studies is that sequence-specific RNA recognition by a single protein may be of considerably less importance in sequence-specific RNA processing reactions like polyadenylation than are cooperative interactions between protein factors that allow limited and variable RNA sequence content to be used effectively in defining specific RNA signals.

ACKNOWLEDGMENTS

We thank R. Tacke for providing the SELEX template and primers and for helpful discussions, L. Ge for technical assistance, and M. Riley for preparing the manuscript.

This work was supported by National Institutes of Health grant GM28983.

REFERENCES

1. **Bardwell, V. J., M. Wickens, S. Bienroth, W. Keller, B. S. Sproat, and A. I. Lamond.** 1991. Site-directed ribose methylation identifies 2′-OH groups in polyadenylation substrates critical for AAUAAA recognition and poly(A) addition. Cell **65:**125–133.
2. **Bienroth, S., E. Wahle, C. Sulter-Crazzolara, and W. Keller.** 1991. Purification of the cleavage and polyadenylation factor involved in the 3′-processing of mRNA precursors. J. Biol. Chem. **266:**19768–19776.
3. **Chanfreau, G., S. M. Noble, and C. Guthrie.** 1996. Essential yeast protein with unexpected similarity to subunits of mammalian cleavage and polyadenylation specificity factor (CPSF). Science **274:**1511–1514.
4. **Chen, J., and C. Moore.** 1992. Separation of factors required for cleavage and polyadenylation of yeast pre-mRNA. Mol. Cell. Biol. **12:**3470–3481.
5. **Chou, Z.-F., F. Chen, and J. Wilusz.** 1994. Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals. Nucleic Acids Res. **22:**2525–2531.
6. **Christofori, G., and W. Keller.** 1988. 3′ cleavage and polyadenylation of mRNA precursors in vitro requires a poly(A) polymerase, a cleavage factor, and a snRNP. Cell **54:**875–889.
7. **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.

8. **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.
9. **Gilmartin, G. M., E. S. Fleming, J. Oetjen, and B. R. Graveley.** 1995. CPSF recognition of an HIV-1 mRNA 3′-processing enhancer: multiple sequence contacts involved in poly(A) site definition. Genes Dev. **9:**72–83.
10. **Guo, Z., and F. Sherman.** 1996. 3′-end-forming signals of yeast mRNA. Trends Biochem. Sci. **21:**477–481.
11. **Jenny, A., H.-P. Hauri, and W. Keller.** 1994. Characterization of cleavage and polyadenylation specificity factor and cloning of its 100-kilodalton subunit. Mol. Cell. Biol. **14:**8183–8190.
12. **Jenny, A., and W. Keller.** 1995. Cloning of cDNAs encoding the 160kDa subunit of the bovine cleavage and polyadenylation specificity factor. Nucleic Acids Res. **23:**2629–2635.
13. **Jenny, A., L. Minvielle-Sebastia, P. J. Preker, and W. Keller.** 1996. Sequence similarity between the 73-kilodalton protein of mammalian CPSF and a subunit of yeast polyadenylation factor I. Science **274:**1514–1517.
14. **Keller, W.** 1995. No end yet to messenger RNA 3′ processing! Cell **81:**829–832.
15. **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 AAUAAA. EMBO J. **10:**4241–4249.
16. **Lutz, C. S., and J. C. Alwine.** 1994. Direct interaction of the U1 snRNP-A protein with the upstream efficiency element of the SV40 late polyadenylation signal. Genes Dev. **8:**576–586.
17. **Lutz, C. S., K. G. K. Murthy, N. Schek, J. P. O'Connor, J. L. Manley, and J. C. Alwine.** 1996. Interaction between the U1 snRNP-A protein and the 160kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. Genes Dev. **10:**325–337.
18. **MacDonald, C. C., J. Wilusz, and T. Shenk.** 1994. The 64-kDa subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol. Cell. Biol. **14:**6647–6654.
19. **Manley, J. L.** 1995. A complex protein assembly catalyzes polyadenylation of mRNA precursors. Curr. Opin. Genet. Dev. **5:**222–228.
20. **Manley, J. L., and Y. Takagaki.** 1996. The end of the message—another link between yeast and mammals. Science **274:**1481–1482.
21. **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.
22. **Minvielle-Sebastia, L., B. Winsor, N. Bonneaud, and F. Lacroute.** 1991. Mutations in the yeast *RNA14* and *RNA15* genes result in an abnormal mRNA decay rate; sequence analysis reveals an RNA-binding domain in the RNA15 protein. Mol. Cell. Biol. **11:**3075–3087.
23. **Minvielle-Sebastia, L., P. J. Preker, and W. Keller.** 1994. RNA14 and RNA15 proteins as components of a yeast pre-mRNA 3′-end processing factor. Science **266:**1702–1705.
24. **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.
25. **Murthy, K. G. K., and J. L. Manley.** 1995. The 160kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3′-end formation. Genes Dev. **9:**2672–2683.
26. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
27. **Proudfoot, N.** 1991. Poly(A) signals. Cell **64:**671–674.
28. **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.
29. **Ryner, L. C., Y. Takagaki, and J. L. Manley.** 1989. Multiple forms of poly(A) polymerases purified from HeLa cells function in specific mRNA 3′-end formation. Mol. Cell. Biol. **9:**4229–4238.
30. **Skolnik-David, H., C. L. Moore, and P. A. Sharp.** 1987. Electrophoretic separation of polyadenylation-specific complexes. Genes Dev. **1:**672–682.

30a. **Smith, D. B., K. M. Davern, P. G. Board, W. U. Tiu, E. G. Garcia, and G. F. Mitchell.** 1986. Mr 26,000 antigen of *Schistosoma japonicum* recognized by resistant WEHI 129/J mice is a parasite glutathione S-transferase. Proc. Natl. Acad. Sci. USA **83:**8703–8707.

31. **Smith, D. B., and K. S. Johnson.** 1988. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. Gene **67:**31–40.
32. **Stumpf, G., and H. Domdey.** 1996. Dependence of yeast pre-mRNA 3′-end processing on CFT1: a sequence homolog of the mammalian AAUAAA binding factor. Science **274:**1517–1520.
33. **Tacke, R., and J. L. Manley.** 1995. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. EMBO J. **14:**3540–3551.
34. **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.
35. **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.
36. **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.
37. **Takagaki, Y., and J. L. Manley.** 1992. A human polyadenylation factor is a G protein β-subunit homologue. J. Biol. Chem. **267:**23471–23474.
38. **Takagaki, Y., and J. L. Manley.** 1994. A polyadenylation factor subunit is the human homologue of the *Drosophila suppressor of forked* protein. Nature **372:**471–474.
39. **Takagaki, Y., R. L. Seipelt, M. L. Peterson, and J. L. Manley.** 1996. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell **87:**941–952.
40. **Tsai, D. E., D. S. Harper, and J. D. Keene.** 1991. U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts. Nucleic Acids Res. **19:**4931–4936.
41. **Tuerk, C., and L. Gold.** 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249:**505–510.
42. **Wahle, E., and W. Keller.** 1996. The biochemistry of polyadenylation. Trends Biochem. Sci. **21:**247–250.
43. **Wilusz, J., and T. Shenk.** 1988. A 64kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. Cell **52:**221–228.
44. **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.
45. **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.