THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 272, No. 42, Issue of October 17, pp. 26769–26779, 1997

Printed in U.S.A.

RNA Ligands Selected by Cleavage Stimulation Factor Contain Distinct Sequence Motifs That Function as Downstream Elements in 3′-End Processing of Pre-mRNA*

(Received for publication, May 16, 1997, and August 13, 1997)

Katrin Beyer, Thomas Dandekar‡, and Walter Keller§

From the Department of Cell Biology, Biozentrum of the University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland and the ‡European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, D-69117 Heidelberg, Federal Republic of Germany

Critical events in 3′-end processing of pre-mRNA are the recognition of the AAUAAA polyadenylation signal by cleavage and polyadenylation specificity factor (CPSF) and the binding of cleavage stimulation factor (CstF) via its 64-kDa subunit to the downstream element. The stability of this CPSF·CstF·RNA complex is thought to determine the efficiency of 3′-end processing. Since downstream elements reveal high sequence variability, *in vitro* selection experiments with highly purified CstF were performed to investigate the sequence requirements for CstF-RNA interaction. CstF was purified from calf thymus and from HeLa cells. Surprisingly, calf thymus CstF contained an additional, novel form of the 64-kDa subunit with a molecular mass of 70 kDa. RNA ligands selected by HeLa and calf thymus CstF contained three highly conserved sequence elements as follows: element 1 (AUGCGUUCCUCGUCC) and two closely related elements, element 2a (YGUGUYN_{0-4}UUYAYUGYGU) and element 2b (UUGYUN_{0-4}AUUUACU(U/G)N_{0-2}YCU). All selected sequences tested functioned as downstream elements in 3′-end processing *in vitro*. A computer survey of the EMBL data library revealed significant homologies to all selected elements in naturally occurring 3′-untranslated regions. The majority of element 2a homologies was found downstream of coding sequences. Therefore, we postulate that this element represents a novel consensus sequence for downstream elements in 3′-end processing of pre-mRNA.

The primary transcripts (pre-mRNAs) of a eukaryotic cell undergo several different maturation steps to become fully functional messenger RNAs (mRNAs). One of these maturation events is the 3′-end processing reaction, during which the pre-mRNA receives a new 3′-end that is in almost all cases a poly(A) tail. First, the pre-mRNA is endonucleolytically cleaved at the polyadenylation site (poly(A) site). In a second tightly coupled event, the polyadenylation reaction, approximately 250 adenosine residues are added to the upstream cleavage product, whereas the downstream fragment is rapidly degraded (for reviews, see Refs. 1–6).

In vivo and *in vitro* studies have revealed a requirement for distinct sequence elements for 3′-end processing of pre-mRNAs, a highly conserved AAUAAA sequence located upstream of the poly(A) site and so-called downstream elements. Moreover, several sequences located upstream of the AAUAAA signal have been shown to enhance the cleavage reaction (for reviews, see Refs. 3, 4, 7).

Downstream elements show a high sequence variability and many different motifs have been proposed to be involved in downstream element function as follows: YGUGUUYY[1] (8), GUGUUG (9), CAYUG (10), AGGUUUUUU (11), UCCUGU (12), or simply UGU clusters (13). Recently, it was shown that a UUUUU element located 6–25 nucleotides downstream of the AAUAAA sequence is sufficient to confer cleavage activity to a substrate whose natural downstream region has been completely deleted (14). Due to the abundance of uracil and guanine residues in these motifs, downstream elements are usually referred to as U- or G/U-rich elements.

One of the best analyzed downstream regions is that of the SV40 late pre-mRNA. Although it was not possible to identify single nucleotides that are essential for poly(A) site function (15), a deletion of about 20 nucleotides downstream of the poly(A) site inhibits 3′-end processing (16–18). The SV40 late downstream element consists of two parts. Each part alone allowed efficient processing when the other part was substituted with unrelated polylinker sequence. Only the substitution of both parts together inhibited cleavage of the SV40 late pre-mRNA (17). Other bipartite downstream elements were identified in the β-globin genes of rabbit (19) and mouse (20).

It has also been demonstrated that the distance between the AAUAAA signal and the downstream element is critical. Moving the downstream element further downstream can not only abolish cleavage but can also shift the cleavage site (14, 19, 21–25).

To date, six factors involved in the cleavage and polyadenylation reactions have been identified as follows: cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factors I_m and II_m (CF I_m, CF II_m), poly(A) polymerase (PAP), and poly(A) binding protein II (for reviews, see Refs. 1, 3, 4). Most of these factors have been purified, and several have been cloned. Interestingly, many homologs to the mammalian 3′-end processing components have been found in yeast suggesting a conserved mechanism in lower and higher eukaryotes (for reviews, see Refs. 2 and 26).

The recognition of the AAUAAA sequence by CPSF is

* This work was supported by grants from the Kantons of Basel and the Swiss National Science Foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed. Tel.: 41-61-267 2060; Fax: 41-61-267 2079; E-mail: Keller2@ubaclu.unibas.ch.

[1] The abbreviations used are: Y, pyrimidine; CPSF, cleavage and polyadenylation specificity factor; CstF, cleavage stimulation factor; CF I_m and CF II_m, mammalian cleavage factor I and cleavage factor II; PAP, poly(A) polymerase; SELEX, systematic evolution of ligands by exponential enrichment (38); 3′-UTRs, 3′-untranslated regions; FPLC, fast protein liquid chromatography; RBD, RNA binding domain; PCR, polymerase chain reaction; nt, nucleotide; DTT, dithiothreitol; Tricine, *N*-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine.

thought to be the first step in the formation of a 3′-end processing complex. This initial complex is stabilized by the subsequent binding of CstF to the downstream element (27–29). The stability of this commitment or ternary complex correlates with the efficiency of poly(A) site usage (29). CF I_m, CF II_m, and PAP then join to form a fully active 3′-end processing complex.

CstF consists of three polypeptides with molecular masses of 50, 64, and 77 kDa (28, 30), all of which have been cloned (31–33). The 64-kDa subunit interacts with the downstream element of pre-mRNAs (28, 30, 31, 34).

A correlation between CstF activity and the usage of different poly(A) sites was observed during the adenoviral life cycle (35) and mouse B-cell development (36). It has been demonstrated that overexpression of the 64-kDa subunit of CstF in stably transformed B-cells induces the switch from the membrane-bound to the secreted form of immunoglobulins via alternative polyadenylation (37). These results demonstrate that CstF plays a critical role in 3′-end processing.

Since the sequence requirements for CstF-RNA interaction have only been poorly characterized, we performed *in vitro* selection experiments (SELEX, Ref. 38) with CstF purified from calf thymus whole cell extracts and HeLa cell nuclear extracts. Interestingly, CstF purified from calf thymus contained an additional polypeptide with a molecular mass of 70 kDa, which represents a novel form of the 64-kDa subunit. CstF preferentially selected highly conserved sequence elements rather than guanine- and/or uracil-rich sequences *per se*. The selected sequences functioned as downstream elements in 3′-end processing *in vitro*, and homologies to them were found in natural 3′-untranslated regions (3′-UTRs) of many genes, suggesting a role of these sequences as downstream elements *in vivo*.

EXPERIMENTAL PROCEDURES

Materials—Macroprep Q resin was purchased from Bio-Rad; Blue Sepharose was prepared as described previously (39). All other column resins and prepacked FPLC columns were from Pharmacia Biotech Inc., as well as RNAguard and m^7GpppG. Phenylmethylsulfonyl fluoride was purchased from Serva, leupeptin hemisulfate and Nonidet P-40 from Fluka, pepstatin from Bachem, and ammonium sulfate from Life Technologies Inc. All restriction enzymes, Moloney murine leukemia virus reverse transcriptase, and polynucleotide kinase were from New England Biolabs; creatine kinase, creatine phosphate, calf intestine alkaline phosphatase, Klenow enzyme, and SP6 RNA polymerase were from Boehringer Mannheim. T7 RNA polymerase was purchased from Stratagene, and *Taq* DNA polymerase (AmpliTaq) was from Perkin-Elmer. DNA sequencing was performed with Sequenase version 2.0 (United States Biochemical Corp.). Cordycepin 5′-triphosphate (3′-dATP), dNTPs, and NTPs were from Boehringer Mannheim; all radioactively labeled NTPs and dNTPs were from Amersham Corp. Polyvinyl alcohol was purchased from Sigma, and dithiothreitol (DTT) was from GERBU Biotechnik GmbH.

Purification of Cleavage Stimulation Factor CstF—Whole cell extract from 2 kg of calf thymus was applied to two DEAE-Sepharose fast flow columns as described previously (39). The flow-throughs were pooled and precipitated with ammonium sulfate (50% saturation). Consecutive backwashes with ammonium sulfate were performed (45, 25, and 20% saturation). The 20% ammonium sulfate pellet (10.7 g of protein) was dialyzed and applied to Blue Sepharose columns. CstF activity was further purified by heparin-Sepharose, Macroprep Q, Mono Q FPLC, Mono S FPLC, Superose 6 FPLC, and poly(U)-Sepharose chromatography. The final poly(U)-Sepharose column (1.5 ml) was equilibrated with 35 ml of 0.3 M KCl in buffer G (50 mM Tris-HCl (pH 7.9), 0.5 mM EDTA, 10% v/v glycerol, 0.02% v/v Nonidet P-40, 0.5 mM DTT, 0.5 mM phenylmethylsulfonyl fluoride, 0.4 μg/ml leupeptin, 0.7 μg/ml pepstatin) and developed with a 20-ml gradient from 0.3 to 2 M KCl. CstF activity eluted around 1 M KCl, and the fractions were pooled and concentrated (Centricon-30, Amicon) to a final protein concentration of 48 μg/ml. This fraction (Fig. 1*A, lane 1*) was used for selection experiments.

For the first purification of HeLa CstF, nuclear extracts (5) were prepared from 8.4×10^{10} HeLa cells (2.6 g of protein). The DEAE-Sepharose flow-through was precipitated with ammonium sulfate (80% saturation), and no further backwashes were performed. Macroprep Q, Mono S, and Superose 6 columns were omitted. Single fractions of the final poly(U)-Sepharose column that contained CstF activity were dialyzed against 20 mM KCl in buffer G. One fraction was used for the selection experiments (Fig. 1*A, lane 2*). The protein concentration of this fraction was 32 μg/ml. A second purification was from 5.8×10^{10} HeLa cells (approximately 2 g of protein). Nonidet P-40 was omitted; the DEAE-Sepharose flow-through was not precipitated with ammonium sulfate. The poly(U)-Sepharose column was loaded at a salt concentration of 0.25 M KCl and developed with a gradient (25 ml) from 0.25 to 2 M KCl. CstF containing fractions were dialyzed against 20 mM KCl in buffer G. The protein concentration of the fraction used for selection experiments (Fig. 1*A, lane 3*) was 64 μg/ml.

Purification of Other 3′-End Processing Factors—CPSF was purified from calf thymus (39), and recombinant bovine PAP was prepared as described previously (40). Crude fractions of CF I_m and II_m were prepared as follows: HeLa nuclear extracts were diluted with buffer G to 75 mM KCl and applied to a DEAE-Sepharose fast flow column equilibrated with 75 mM KCl in buffer G. The column was developed with a gradient (10 column volumes) from 75 to 500 mM KCl. Fractions containing CF I_m/II_m activity were pooled and loaded directly onto an 8-ml Mono Q FPLC column. The column was developed with a gradient (25 column volumes) from 100 to 500 mM KCl in buffer G. CF I_m/II_m activity eluted between 250 and 300 mM KCl. These fractions were dialyzed against 100 mM KCl in buffer G and used for cleavage reactions.

Selection of RNA Ligands—DNA oligonucleotides were used to transcribe RNA substrates for the first round of the SELEX procedure (38). Oligo 1 (5′ TAGGCTAGGATCCATCTTGT(N_{20})ATCGTTCGTGAGCTCGTCCCTATAGTGAGTCGTATTACGCG 3′) contained a *Bam*HI restriction site in the 5′ part, a T7 RNA polymerase promoter sequence (underlined) and a *Sac*I restriction site in the 3′ part, and encoded an RNA of 60 nt (5′ GGGACGAGCUCACGAACGAU(N_{20})ACAAGAUGGAUCCUAGCCUA 3′). 4 pmol of both oligo 1 and oligo 2 (5′ CGCGTAATACGACTCACTATAGGG 3′; complementary to the T7 RNA polymerase promoter sequence) were annealed. Transcriptions were performed as recommended by the manufacturer at 37 °C for 1 h, and transcripts were gel-purified. Filter binding reactions (20 μl) contained 0.2 mM DTT, 0.01% v/v Nonidet P-40, 20 mM creatine phosphate, 0.5 mM ATP, 1.5 mM $MgCl_2$, CstF as indicated and were incubated at 30 °C for 30 min. The reaction mixtures were filtered under vacuum through BA 83 nitrocellulose filters (Schleicher und Schüll) that had been equilibrated with buffer W (50 mM KCl, 50 mM Tris-HCl (pH 7.9), 1.5 mM $MgCl_2$, 0.5 mM DTT) and saturated with 20 μg of *Escherichia coli* total RNA. Filters were washed with 4 ml of ice-cold buffer W. The selected RNAs were eluted from the nitrocellulose filters with 350 μl of urea elution buffer (41). The eluate was extracted with phenol/chloroform and ethanol-precipitated, and the RNA pellet was resuspended in 10 μl of H_2O containing 50 pmol of oligo 3 (5′ TAGGCTAGGATCCATCTTGT 3′). After annealing of oligo 3 to the RNA, reverse transcription was performed in a volume of 30 μl in the presence of 25 units of RNAguard and 15 units of Moloney murine leukemia virus reverse transcriptase as recommended by the manufacturer. After 45 min at 37 °C, 70 μl of H_2O was added, and the DNA was extracted with phenol/chloroform, ethanol-precipitated, and resuspended in 30 μl of H_2O. One-third of this sample was amplified by PCR in a final volume of 50 μl in the presence of 10 mM Tricine (pH 8.4), 50 mM KCl, 0.01% w/v gelatin, 1.5 mM $MgCl_2$, 2.5 units of *Taq* DNA polymerase, 0.4 mM of each dNTP and 50 pmol of both oligo 3 and 4 (5′ CGCGTAATACGACTCACTATAGGGACGAGCTCACGAACGAT 3′). PCRs were performed in a HYBAID reactor (Biotechnology LTD; 30 cycles: 15 s 94 °C, 30 s 51 °C, 30 s 72 °C; 1 cycle 5 min 72 °C). The DNA was gel-purified, phenol/chloroform-extracted, ethanol-precipitated, and resuspended in 10 μl of H_2O. 2–5 μl of this DNA were used either to transcribe RNA for the next round of selection or was subcloned. The following selection conditions were applied as follows: the first round used 4 pmol of CstF (0.2 μM), and subsequent rounds used 0.4 pmol (0.02 μM). The RNA concentration varied as follows: round 1, 0.2 μM; rounds 2 and 3, 0.2 μM; round 4, 0.4 μM; rounds 5 and 6, 1 μM; round 7, 2 μM; and round 8, 20 μM. PCR products from the final round of selection were digested with *Bam*HI and *Sac*I and subcloned into Bluescript KS vectors (Stratagene) for sequencing.

RNA Substrates—SV40 wild type RNA was transcribed from the plasmid pSV-L (15). The plasmid pSV-141/-1 (17) is an SV40 late derivative of which the complete downstream region was replaced by a *Xba*I linker and pBR322 sequences. The *Xba*I site located in the polylinker region of the plasmid pSV-141/-1 was deleted by digestion with *Bam*HI and *Sal*I, the recessed 3′-termini were filled with Klenow enzyme. The resulting plasmid (pSVΔ-1) contains a single *Xba*I site immediately downstream of the natural polyadenylation site. DNA

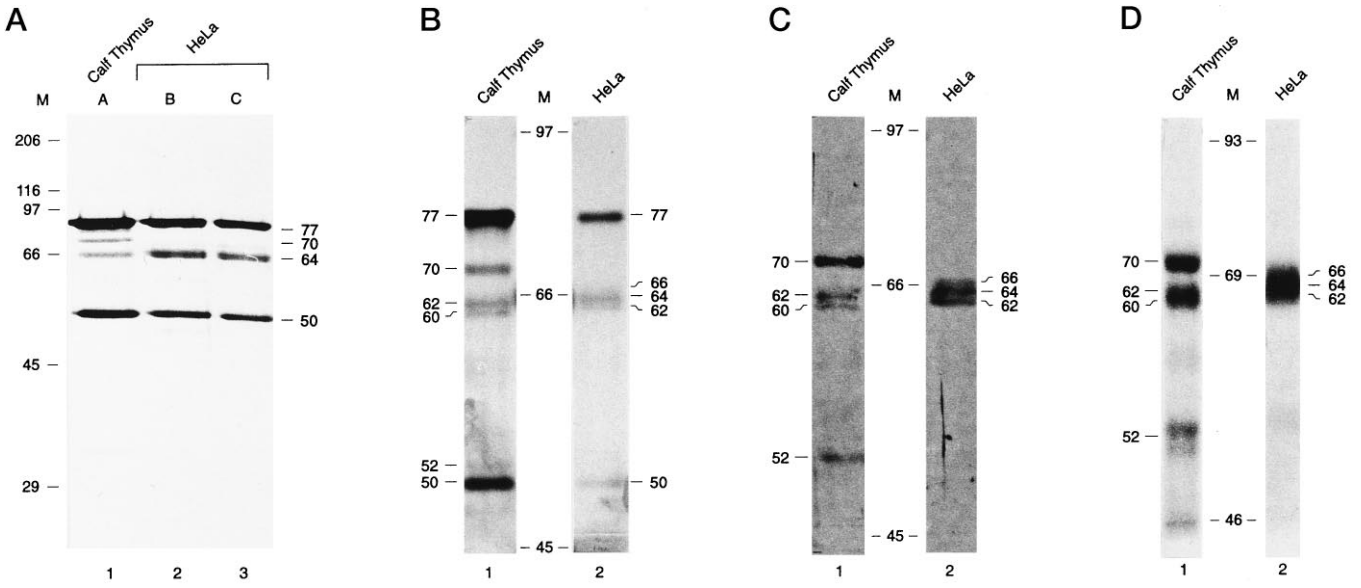

FIG. 1. **Polypeptide composition of calf thymus and HeLa CstF.** *A*, purified CstF derived from either calf thymus whole cell extract (*lane 1*) or HeLa cell nuclear extracts (*lanes 2* and *3*; for details, see "Experimental Procedures") was separated on an SDS-10% polyacrylamide gel and stained with silver. All three CstF preparations were used for the selection of RNA-ligands (*pools A, B,* and *C*; indicated at the *top*). Purified calf thymus and HeLa CstF was separated on an SDS-7.5% polyacrylamide gel and stained with silver (*B*) or used for immunoblot analysis (*C*). Immunodetection was performed with polyclonal antibodies directed against the human 64-kDa subunit of CstF. *D*, UV cross-linking of calf thymus and HeLa CstF to the selected RNA A-1. The molecular masses of size markers (*M*) as well as those of the CstF polypeptides are indicated.

oligonucleotides containing *Xba*I site overhangs (CTAG) at their 5′-ends and encoding sequences of interest in either sense or antisense orientation were annealed and subcloned into the *Xba*I site of pSVΔ-1. The correct insertions were confirmed by sequencing. All pSVΔ-1 derivatives were linearized with *Eco*RI, pSV-L with *Dra*I, and uniformly labeled RNA substrates were obtained by SP6 RNA polymerase transcription and gel purification (42).

Cleavage Assays—Cleavage reactions were performed as described previously (42) with the following modifications: 20 fmol radioactively labeled RNA substrate were used and reactions were incubated for 1 h at 30 °C. In some experiments, 0.5 mM 3′-dATP and 1.5 mM $MgCl_2$ were replaced by 1 mM ATP and 1 mM EDTA. HeLa CstF was titrated in the range of 0 and 10 ng, and crude CF I_m/II_m fractions (10 μl) were used. The cleavage reactions were quantitated with a PhosphorImager 425 (Molecular Dynamics) and IPlab Gel (version 1.5, Signal Analytics Corp.). The calculation of cleavage activity took into account the loss of radioactivity present in the downstream cleavage fragment. To compare the cleavage activities of SVΔ-1 derivatives with SV40 late pre-mRNA, the percentage of cleavage obtained with 4.5 ng of CstF and SV40 was set to 100%. These relative cleavage activities were determined in at least three independent cleavage reactions for each construct (except for SV-B13: two independent experiments), and their averages are presented in Fig. 3.

Immunoblot Analysis—Proteins were separated on an SDS-7.5% polyacrylamide gel, blotted on nitrocellulose, and detected with chemiluminescence staining (ECL kit, Amersham Corp.) as recommended by the manufacturer. The CstF-64 polyclonal antibodies were diluted 1:10000.

UV Cross-linking—100 fmol of CstF and 400 fmol of radioactively labeled RNA (Fig. 2, *A-1*) were incubated in 12.5 μl including 2 mM DTT, 20 mM creatine phosphate, 0.5 mM ATP, 1.5 mM $MgCl_2$, 0.01% Nonidet-P40, 0.1 μg/ml bovine serum albumin for 20 min at room temperature. After UV irradiation (500 kJ, Stratalinker UV1800, Stratagene), 200 ng of RNase were added and reactions were incubated for 30 min at 37 °C before separation on a SDS-7.5% polyacrylamide gel. The gel was fixed (20% 2-propanol, 10% acetic acid), dried, and exposed to Kodak X-Omat AR films.

Computer Surveys—The different consensus elements selected by CstF were translated into specific search programs written in VAX Pascal (43) and used to screen the EMBL data library (44; release 48). First, the programs searched for the presence of the polyadenylation signal AATAAA (no mismatch allowed). If this signal had been found, for each consensus element a search up to 50 nt downstream would be conducted subsequently. Pool 1 was obtained with element 1 demanding ATGCGTT with at least 5 matches and CCTCGTCC directly following. Pool 2a, screened with element 2a, demanded the sequence YGTGTY with at least 4 matches, directly or up to 5 nt later followed by TTYAYTG with at least 4 matches and directly or up to 2 nt later the sequence YGT. Two pool 2b screens with element 2b were performed. The first for pool 2b sequences (pool $2b/T_4$) required the TTGYT with at least 3 matches, directly or up to 5 nt later followed by ATTTTACT(T/G) with at least 3 matches and directly or up to 2 nt later the sequence YCT. The second screen (pool $2b/T_3$) differed from pool $2b/T_4$ in that ATTTACT(T/G) with at least 3 matches was required instead of ATTTTACT(T/G). As a positive control, the pool M was screened for the presence of the consensus sequence for downstream elements YGTGTTYY proposed by McLaughlan *et al.* (8). No minimum number of matches was demanded for this screen.

RESULTS

Polypeptide Composition of Calf Thymus and HeLa CstF—CstF was purified from calf thymus whole cell extracts and twice independently from HeLa cell nuclear extracts (for details, see "Experimental Procedures"). Fractions of the final poly(U)-Sepharose columns used for the selection of RNA ligands (see below) are shown in Fig. 1*A*. HeLa CstF consists of three polypeptides with molecular masses of 50, 64, and 77 kDa (Refs. 27, 28, 30, 45; Fig. 1*A, lanes 2* and *3*). CstF purified from calf thymus also contained the 50- and 77-kDa polypeptides but differed from HeLa CstF in that the 64-kDa subunit was much less abundant and contained an additional polypeptide with a molecular mass of approximately 70 kDa (Fig. 1*A, lane 1*). To investigate whether this 70-kDa subunit represents an alternative form of the 64-kDa polypeptide, calf thymus and HeLa CstF were separated on high resolution SDS-polyacrylamide gels and either stained with silver (Fig. 1*B*) or used for Western blot analysis and immunodetection with polyclonal antibodies raised against the human 64-kDa subunit (Fig. 1*C*). On this gel, the 64-kDa subunit of HeLa CstF emerges as a doublet of 64 and 62 kDa and a significantly less abundant 66-kDa polypeptide (Ref. 30; Fig. 1*B, lane 2*). The 64-kDa subunit of calf thymus CstF is a doublet of 62 and 60 kDa (Fig. 1*B, lane 1*). Furthermore, a less abundant polypeptide with a molecular mass of approximately 52 kDa is visible (Fig. 1*B, lane 1*). Both 64-kDa doublets, the HeLa 66-kDa polypeptide and the calf thymus 52- and 70-kDa polypeptides, were recognized by the polyclonal antibodies (Fig. 1*C*). However, the monoclonal anti-

TABLE I
Nucleotide composition of the starting pool 0 and the selected pools A, B, and C

The substrate RNAs (60 nt) contained 20 random nucleotides in the middle. Eight rounds of selection with increasing stringency were performed. Pool 0 was used for the first round of selection; pool A was selected by calf thymus CstF and pools B and C by two independent preparations of HeLa CstF.

	Pool 0	Pool A	Pool B	Pool C
Apparent K_D (nM)	360	5.5	5.6	5.5
No. inserts analyzed[a]	74	62	53	63
No. different sequences	74	10	22	14
No. nucleotides[b]				
Adenine	3.9	1.7	1.9	2.0
Guanine	4.3	5.3	5.3	4.4
Cytosine	6.7	3.2	5.9	6.5
Uracil	5.1	9.8	6.9	7.1

[a] The template DNA oligonucleotides of pool 0 were amplified by PCR, the RNA pools A, B, and C by reverse-transcriptase-PCR, subcloned into Bluescript KS vectors, and sequenced.

[b] The average numbers for all nucleotides are given for a 20-nt sequence with the length variations of some inserts taken into account.

body 3A7 (30) directed against the human 64-kDa subunit did not recognize the 70-kDa subunit (data not shown). In addition, the 52- and 70-kDa subunits can be as efficiently UV crosslinked to RNA as the other 64-kDa polypeptides (Fig. 1*D*). These results reveal differences in the polypeptide composition of calf thymus and HeLa CstF in respect to the 64-kDa subunit that interacts with the RNA. The precise nature of these differences must await the cloning of cDNA coding for the new subunit.

Selection of RNA Ligands by CstF—To investigate the sequence requirements for CstF-RNA interaction, purified calf thymus and HeLa CstF (Fig. 1*A*) were used for *in vitro* RNA selection experiments (38). The RNA substrates (60 nt) in which the central 20 nucleotides were randomized were subjected to filter binding reactions with pure CstF. Eight rounds of selection with increasing stringency were performed. To ensure that any putative ligand of CstF was selected during the first round of selection, a CstF:RNA ratio of 1:1 was chosen (0.2 μM CstF). During the subsequent steps of selection, the CstF concentration was kept constant (0.02 μM), whereas the RNA concentration was increased progressively from 0.2 to 20 μM corresponding to CstF:RNA ratios from 1:10 to 1:1000, respectively (for details, see "Experimental Procedures"). The RNA pool A was selected by calf thymus CstF (Fig. 1*A, lane 1*) and pools B and C by two independent preparations of CstF from HeLa cells (Fig. 1*A, lanes 2* and *3*). During this procedure, the affinities of the selected RNA pools increased 65-fold (apparent K_D values approximately 5.6 nM; Table I) in comparison to the starting pool 0 (apparent K_D 360 nM).

To analyze the RNA pools, PCR products of the starting pool 0 and the last selection rounds (pools A, B, and C) were subcloned, and 50 clones of each pool were sequenced. Some of these clones contained multiple insertions, so that altogether between 52 and 74 inserts of each pool were analyzed (Table I). No identical sequences were found in pool 0. In contrast, the number of different sequences was drastically reduced in those pools that have been subjected to selection by CstF: only 10 different sequences were found in pool A, 22 in pool B, and 14 in pool C (Table I). The sequences of these inserts are presented in Fig. 2.

The average abundance of the nucleotides in all pools is shown in Table I. A frequency of 5 out of 20 nt corresponds to a random distribution and was expected for all nucleotides in pool 0. This was only true for uracil (5.1) and guanine (4.3) but not for adenine (3.9) and cytosine (6.7). These deviations might be due to the unequal use of nucleotides during DNA-oligonucleotide synthesis. Upon CstF selection, the adenine content was reduced (1.7–2.0), the guanine content was nearly unchanged (4.4–5.3), and the amount of cytosine was only diminished in the calf thymus pool A (3.2). Instead, the selected RNAs were enriched in uracil, which was more significant in pool A (9.8) than in those pools that were selected by HeLa CstF (6.9, 7.1).

As shown in Table I, the number of different sequences in the RNA pools decreased upon CstF selection. Pool A, which was selected by calf thymus CstF, contained mainly two different RNAs (Fig. 2, *A-1*, 45.2% and *A-2*, 33.9%). The most abundant inserts of pool B, selected by HeLa CstF, were B-2 (15.1%), B-1, B-10, and B-15 (13.2% each). Pool C consisted of three prominent sequences (C-1, 36.5%, C-3A, 28.6%, and C-5, 14.3%).

Sequence compilation of all selected RNAs led to the deduction of three different sequence elements, element 1 and two closely related elements 2a and 2b (Fig. 2). Element 1 had the consensus AUGCGUUCCUCGUCC, and each nucleotide of this element was conserved with a frequency of at least 75%, five residues were conserved to 100% (Fig. 2). Nucleotides 2–9 of this element were homologous to the consensus sequence for downstream elements YGUGUUYY (8). With exception of one single RNA (A-22/3: 60% identity), all RNAs of this group were at least 87% identical to the derived consensus sequence.

A second highly conserved element was YGUGU-YN_{0-4}UUYAYUGYGU (Fig. 2, *element 2a*). Each nucleotide was conserved with a frequency of at least 85%, and three residues were conserved to 100% (Fig. 2). Element 2a had a bipartite structure as follows: a GU motif (YGUGUY) in the 5′ part and a pyrimidine-rich part (AY motif, UUYAYUG) containing a highly conserved adenine residue (94% conservation) followed by a second shortened GU motif (YGU). The distances between the first two parts were variable, 65% of all inserts aligned had insertions of 1–4 nt between the 5′ GU motif and the AY motif. Only a few (8%) of the selected sequences also had insertions (1 to 2 nt) between the AY motif and the 3′ YGU. With exception of four selected RNAs (A-7 and B-7, 63% identity; B-5 and C-37, 75% identity), all RNAs were at least 88% identical to the consensus of element 2a.

The third element, element 2b UUGYUN_{0-4}AUUUACU-GN_{0-2}YCU, strongly resembled element 2a but was less conserved (Fig. 2). Only the inserts of pool C shared at least 76% homology to this element; A-11 and B-2 were 59% identical, and B-1 and B-36 were 69% homologous. Element 2b differed from element 2a in that both the 5′-GU motif and the AY motif were slightly altered. Furthermore, 88% had insertions (1 to 2 nt) between the AY motif and the 3′-YCU. Interestingly, the two nucleotides inserted between the AY motif and the last three nucleotides of both elements 2a and 2b were in most cases adenine and cytosine residues, which therefore formed a second, shortened AY motif.

Our results demonstrate that distinct sequence elements rather than random guanine and uracil residues are required for efficient CstF-RNA interaction. Furthermore, different sequence elements were selected to different extents by calf thymus and HeLa CstF; element 1 was present more frequently in those pools that were selected with HeLa CstF (pool B, 34.0%; pool C, 41.3%) than in the pool selected with calf thymus CstF (Fig. 2, pool A, 9.7%), whereas the closely related elements 2a and 2b were the most frequent elements in the calf thymus pool (sum of elements: pool A, 90.3%; pool B, 66.0%; pool C, 58.7%). These findings might indicate that differences in the polypeptide composition of calf thymus and HeLa CstF may be reflected by slightly altered RNA binding properties.

In Vitro 3′-End Processing Reactions with Selected Sequences

as Downstream Elements—Selected element 1 and 2a sequences and minimal versions of both were tested for their ability to restore cleavage activity of the non-functional SV40 late pre-mRNA derivative SVΔ-1, whose natural downstream region had been substituted by an *Xba*I linker and unrelated sequence. DNA oligonucleotides encoding the sequences of interest were inserted into this *Xba*I site (for details, see "Experimental Procedures"). The sequences of the first 49 nt following the AAUAAA hexamer of all RNA substrates used are shown in Fig. 3*A*.

Three substrates (SV-A42, SV-B13, and SV-C1) contained selected element 1 sequences as downstream elements, whereas others contained only the most highly conserved part of element 1 (11-mer element, UGCGUUCCUCG) at different positions of the inserted oligonucleotides (SV-E1P4 and SV-E1P12) or as a duplication (SV-E1d). The sequence of SV-αB14, which carried the selected sequence B-14 in antisense orientation and which was processed as inefficiently as SVΔ-1 in *in vitro* cleavage reactions, was used to embed the 11-mer elements of SV-E1P4 and SV-E1P12.

The processing efficiencies of these constructs were determined in reconstituted *in vitro* cleavage reactions with highly purified PAP, CPSF, and partially purified CF I_m/II_m fractions in non-limiting amounts, whereas highly purified HeLa CstF was titrated between 0 and 10 ng (for details, see "Experimental Procedures"). Typical cleavage reactions are shown in Fig. 3*B*. At least three independent titration experiments were performed for each construct (except SV-B13, two independent experiments), of which the averages relative to SV40 are presented in Fig. 3*C*. The amount of cleavage activity obtained with 4.5 ng of CstF and SV40 was defined as 100%, and the relative cleavage activities at this reference point for all RNAs are summarized in Fig. 3*A*.

All element 1 sequences restored cleavage activity of SVΔ-1 (Fig. 3, *B* and *C, left panel*). SV-A42 and SV-B13 were nearly as efficiently processed as SV40 (75–80%), whereas SV-C1 was less active (55%). Furthermore, the 11-mer element alone was able to function as a downstream element (Fig. 3*C, right panel*). SV-E1P4 was processed with moderate efficiency (45%), whereas SV-E1P12, of which the 11-mer element was shifted further downstream, was processed with high efficiency (80%). Duplication of this 11-mer element (SV-E1d) restored cleavage activity to wild type levels (Fig. 3*A*, *95%*), and furthermore, SV-E1d was the only construct that was cleaved exclusively at the natural poly(A) site. A shift of the cleavage site was observed for all other element 1 constructs as follows: SV-B13, SV-C1, and SV-E1P4 were cleaved 4 nt, SV-A42 7 nt further downstream of the natural poly(A) site. SV-E1P12 was processed at three different sites; the major cleavage site was the natural poly(A) site but additional sites 7 and 12 nt further downstream were efficiently used as well.

To investigate whether element 2a can function as a downstream element in 3′-end processing *in vitro*, RNA substrates containing selected sequences or variants of this element were analyzed as described for element 1. The sequences of all element 2a containing RNA substrates as well as their relative cleavage activities are summarized in Fig. 3*A*. The selected sequences A-1 and B-14 restored cleavage activity of SVΔ-1 with moderate efficiencies (40–55%; Fig. 3, *A* and *D, left panel*) but to the same extent as the element 1 construct SV-E1P4 (55%). This moderate activity is probably due to a non-optimal position of element 2a relative to the AAUAAA signal.

To investigate the requirement of both the GU motif and the AY motif of element 2a for downstream element function, several variants were constructed that contained these motifs embedded into the non-functional SV-αB14 sequence. SV-E2a contained extended, SV-G/A significantly shortened element 2a sequences (Fig. 3*A*). SV-E2a was cleaved more efficiently (60%) than the minimal element 2a substrate SV-G/A (45%). But the minimal substrate SV-G/A was still processed to the same extent as SV-A1 (40%), which contains the complete selected sequence A-1 (Fig. 3, *A* and *D, left panel*). Substrates containing a deletion of either of these motifs (SV-G/0 and SV-0/A) were only poorly processed (15–20%; Fig. 3, *A* and *D, right panel*). This indicates that both motifs are required for optimal CstF-RNA interaction.

To investigate whether the GU and AY motif are functionally equivalent, RNAs were created that contained these motifs in inverted order or either of them duplicated. A switch of the positions of the minimal GU and AY motifs (SV-A/G, 30%; Fig. 3, *A* and *D, right panel*) as well as a duplication of the minimal AY motifs (SV-A/A, 20%) led to a further reduction of the cleavage activity in comparison to SV-G/A. In contrast, SV-G/G containing two minimal GU motifs was processed as efficiently as SV-G/A and reached nearly wild type activity at higher CstF concentrations (Fig. 3, *A* and *D, right panel*). This indicates that the GU motif can substitute for the AY motif, but not vice versa, and suggests that the GU motif is the more important part of element 2a.

Taken together, these results demonstrate that the sequences selected by CstF as well as shortened versions are able to function as downstream elements in 3′-end processing of pre-mRNA, although to different extents. Also, the fact that the selected sequences functioned as downstream elements in the absence of their constant flanking regions indicates that the flanking sequences played no essential role in CstF binding during the SELEX procedure.

Computer Survey of a Data Library for Homologies to the Selected Elements—To investigate whether the selected sequence elements can be found in 3′-untranslated regions of genes and thus might also function as downstream elements *in vivo*, a computer survey was performed. The EMBL data library was screened with appropriate programs that searched for the presence of a perfect match to the polyadenylation signal AATAAA. This pool (pool V) comprised 45,889 vertebrate and viral sequences, which were subsequently screened for the presence of either of the selected elements up to 50 nt downstream of the AATAAA signal. Pool 1 was obtained with element 1 (ATGCGTTCCTCGTCC; for details, see "Experimental Procedures") and pool 2a with element 2a allowing a second gap in the 3′ part of the motif (TGTGTYN_{0-5}TTYAYTGN_{0-2}YGT). Two screens were performed with element 2b. Pool 2b/T_3 was obtained with the short version (TTGYTN_{0-5}ATTTACT(T/G)N_{0-2}YCT), and pool 2b/T_4 was obtained with the longer variant of element 2b (TTGYTN_{0-5}ATTTTACT(T/G)N_{0-2}YCT). The gaps between the first and the second parts of elements 2a and 2b were increased to five nucleotides according to alignments of these motifs with already identified downstream regions (Refs. 8 and 34; data not shown). As a control, the pool M was generated by screening for the presence of the consensus sequence for downstream elements YGTGTTYY (8).

The number of sequences obtained for the different pools are presented in Table II in respect to the degree of homology to the requested element. The majority of pool V sequences did not fulfill the minimum requirement for the pool 1 screen that demanded at least 5 matches to the first part of element 1 (ATGCGTT). Furthermore, only sequences with not more than 13 matches (87% identity) were found in this pool. In contrast, the pools 2a, 2b, and M contained fewer sequences that did not fulfill the minimum requirements for the distinct screens, and sequences with 100% identity were found. These discrepancies

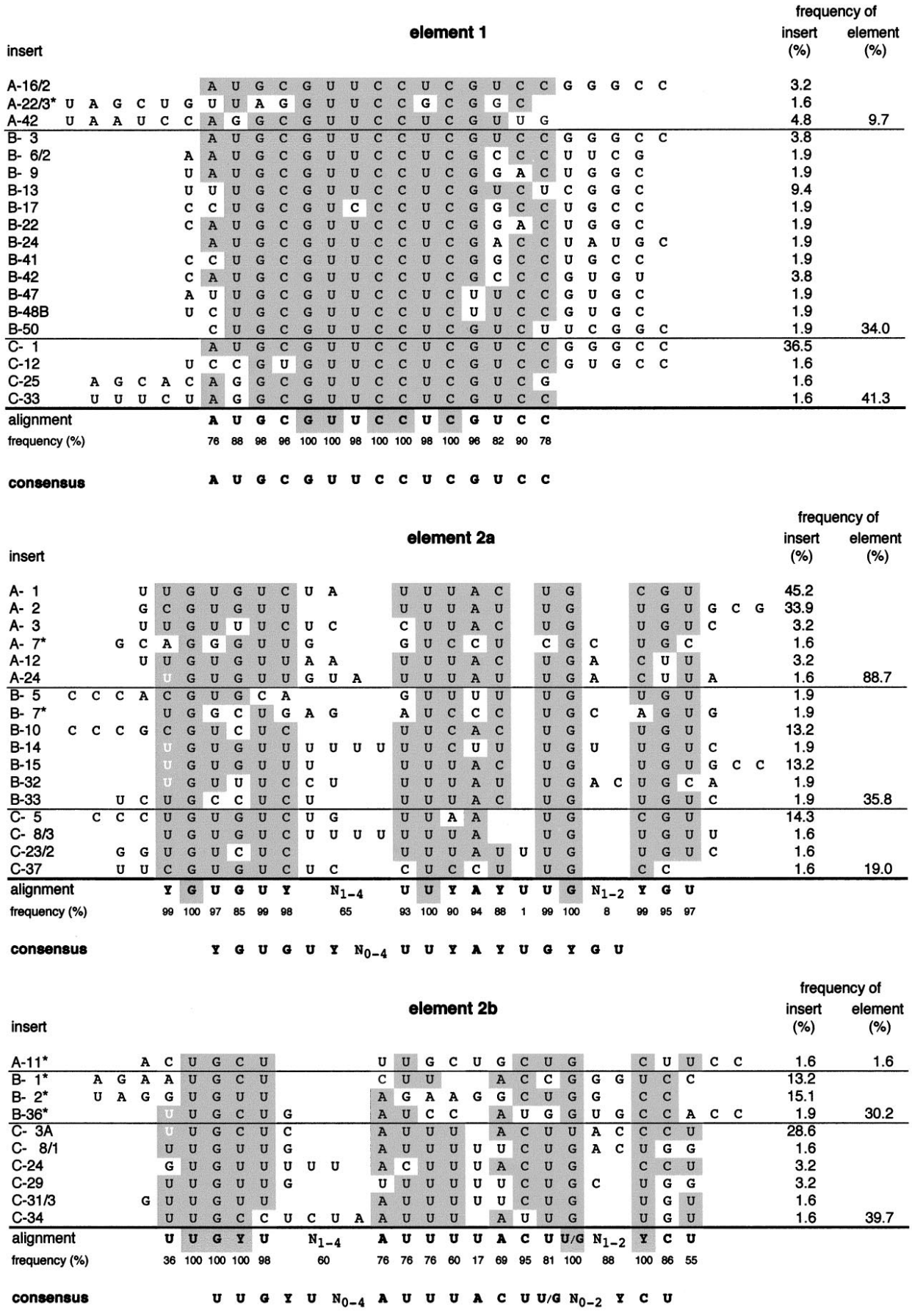

FIG. 2. **Distinct sequence elements are selected by CstF.** All sequences selected by calf thymus CstF (named *A-x*) or HeLa CstF (named *B-x* and *C-x*) are shown, and their frequencies in the corresponding pools are indicated on the *right* (frequency of insert). Some sequence alignments included the last uracil (*white letter*) of the 5′ constant region of the template RNA. Those nucleotides that led to the consensus sequences are

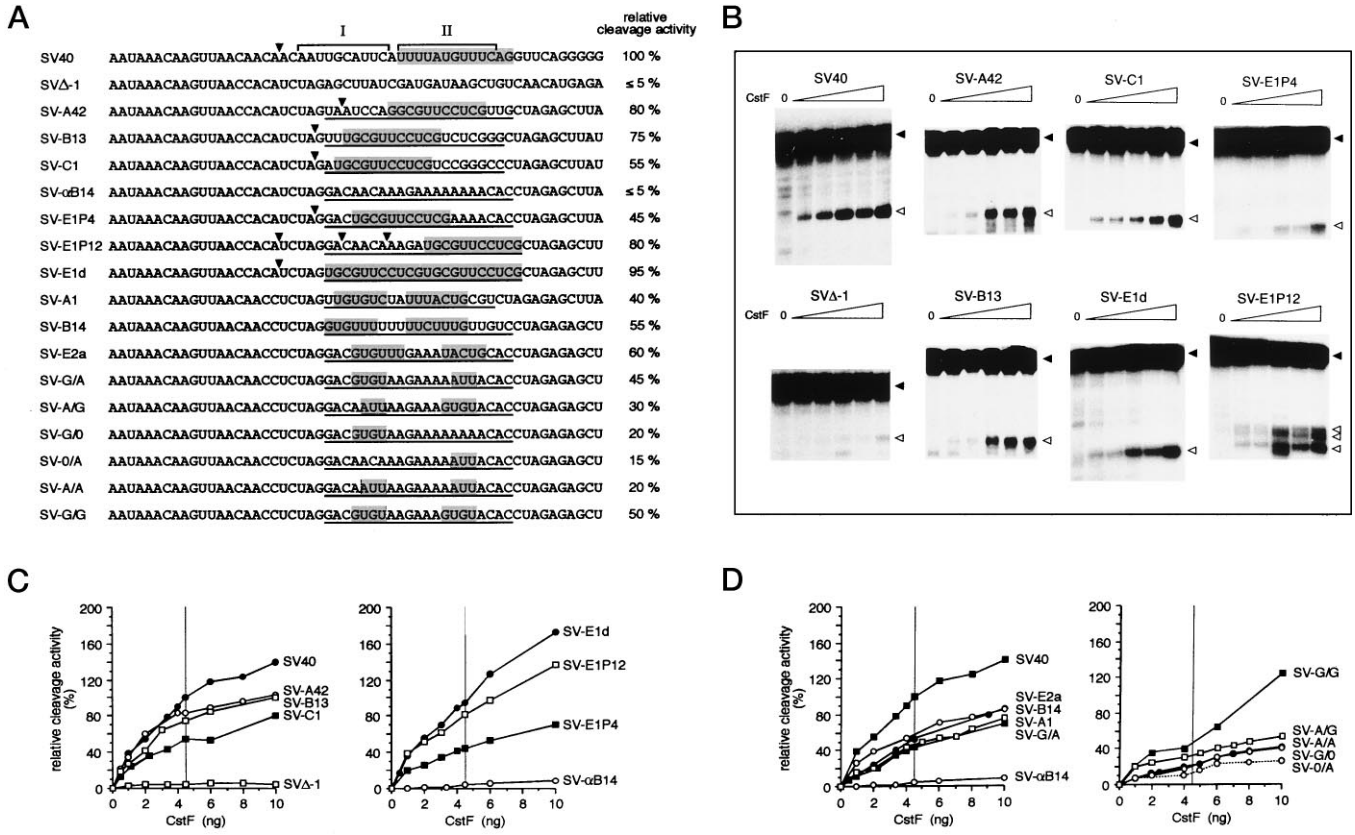

FIG. 3. **The selected elements function as downstream element in 3′-end processing *in vitro*.** Either selected sequences or artificial constructs carrying shortened element 1 or element 2a sequences were tested for their ability to restore 3′-end processing of the cleavage-deficient pre-mRNA SVΔ-1, which lacks its natural downstream element (for details, see "Experimental Procedures"). *A,* RNA sequences of SV40 late, SVΔ-1, and derivatives. The bipartite structure of the SV40 downstream element is indicated by *brackets* and *numbers above* the sequence (17), the CstF-binding site is *screened* in *gray* (34). The sequences inserted into SVΔ-1 are *underlined*, and the minimal element 1 (11-mer element, UGCGUUCCUCG) and the GU and AY motifs of element 2a are *screened* in *gray*. *SV-A42, SV-B13,* and *SV-C1* carry the selected sequences A-42, B-13, and C-1. *SV-αB14* contains the selected sequence B-14 in antisense orientation and was used to embed the 11-mer element and shortened element 2a sequences. *SV-E1P4* and *SV-E1P12* contain the 11-mer element at positions 4 and 12 of the inserted sequences, respectively. SV-E1d encodes a duplicated 11-mer element. SV-E2a carries extended GU and AU motifs, *SV-G/A* contains a minimal GU motif (GUGU) and a minimal AY motif (AUU). Other RNA substrates contain these minimal motifs in different combinations. The experimentally determined positions of the cleavage sites of some substrate RNAs are indicated by *arrowheads*. The average cleavage activities are given as relative cleavage activities in comparison to SV40 late. *B*, cleavage reactions. HeLa CstF was titrated between 0 and 10 ng, whereas the complementing factors (CPSF, PAP, and CF I_m/II_m) were present in non-limiting amounts (for details, see "Experimental Procedures"). *Black arrowheads* indicate the precursor RNA, and *open arrowheads* indicate the 5′ cleavage products. Quantitation of cleavage activities of substrates carry either element 1 (*C*) or element 2a sequences (*D*). The cleavage reactions were quantitated as described under "Experimental Procedures." The value obtained for SV40 with 4.5 ng of CstF was set at 100%, and this reference point is indicated by a *vertical line*. The average of at least three independent experiments is presented (except for SV-B13), and the relative cleavage activities of all constructs are summarized in *A*.

are most likely due to the fact that element 1 is strictly conserved, whereas both elements 2a and 2b are degenerate due to the pyrimidines and the gaps in their consensus sequences.

To analyze whether the sequences identified by the computer survey might be putative downstream elements *in vivo*, the locations of these sequences in the genes were investigated. Sequences of pool 1 with at least 10 matches, of pool 2a with at least 14 matches, of both pools 2b (T_3 and T_4) with at least 15 matches, and 175 sequences of pool M were analyzed. After the elimination of all non-vertebrate virus sequences, duplications or sequences that did not contain any coding sequence, 322 sequences of pool 1, 179 of pool 2a, 61 of pool 2b, and 68 of pool M were analyzed in detail. As shown in Table III, 32% of pool 1 sequences contained the homology to element 1 inside the coding sequence, 45% downstream of it, and 23% were found in introns. In contrast, in only 13% of pool 2a, 8% of pool 2b, and 12% of pool M sequences were the distinct motifs present within the coding sequences. The homologies in pool 2b and pool M sequences were located either in introns (43 and 31%, respectively) or in 3′-UTRs (49 and 57%, respectively), and pool 2a sequences were found in 70% of the cases downstream of the coding sequence. The majority of all elements downstream of the coding sequence were in the context of the first AATAAA signal. Several examples for pool 1, 2a and 2b sequences that contained the distinct motifs downstream of the coding sequence are presented in Fig. 4. Taken together, homologies to all selected elements can be found in the 3′-UTRs of several genes and probably function as downstream elements in 3′-end processing *in vivo*.

DISCUSSION

Calf Thymus and HeLa CstF Contain Different RNA-binding Subunits—CstF was purified to homogeneity from calf thymus whole cell extract and HeLa cell nuclear extracts. Interestingly,

screened in *gray*. The alignments are presented for each consensus element, and the frequency of every nucleotide is given in percentage, residues conserved to 100% are *screened* in *gray*. The abundance of each insert in the corresponding pool was taken into account. The derived consensus sequences are shown for each element at the bottom. The frequency of each element in the different pools is shown on the *right* (frequency of element). Those inserts that shared less than 75% identity to the consensus sequences are indicated by *asterisks*.

TABLE II

Frequency of homologies between the elements selected by CstF and vertebrate and viral sequences

The EMBL data library (release 48) was screened for sequences containing the polyadenylation signal AATAAA. This pool comprised 45,889 vertebrate and viral sequences, which were subsequently screened for the presence of either of the sequence elements selected by CstF (Fig. 2) up to 50 nt downstream of the AATAAA signal (for details, see "Experimental Procedures"). NA, not applicable.

Matches	Pool 1[b]	Pool 2a[b]	Pool 2b/T_4[b]	Pool 2b/T_3[b]	Pool M[b]
0[c]	26,928	5,729	2,313	2,313	394
≤5	1,668	1,476	2,770	2,539	19,867
6	4,070	1,293	1,805	1,964	17,376
7	5,412	4,435	1,912	2,318	7,458
8	4,723	1,316	2,850	3,162	794
9	2,234	3,625	5,140	6,241	NA
10	721	6,806	7,957	9,667	NA
11	114	9,432	9,615	10,031	NA
12	18	7,781	7,102	5,698	NA
13	1	3,310	3,332	1,671	NA
14	0	634	952	257	NA
15	0	50	123	24	NA
16	NA	2	11	4	NA
17	NA	NA	4	NA	NA

[a] Number of nucleotides identical to the requested sequence elements.

[b] Pool 1 was obtained with element 1 (ATGCGTTCCTCGTCC), pool 2a with element 2a (YGTGTYN$_{0-5}$TTYACTGN$_{0-2}$YGT), pool 2a/T_4 with the long version of element 2b (TTGYTN$_{0-5}$ATTTTACT(T/G)N$_{0-2}$YCT), pool 2a/T_3 with the short version of element 2b (TTGYTN$_{0-5}$ATTTACT(T/G)N$_{0-2}$YCT), and pool M with the consensus sequence for downstream elements (YGTGTTYY, Ref. 8).

[c] Except for pool M, the sequence elements used for the data library screens demanded a minimum number of 3 to 5 matches to the first part of the motifs. Sequences that did not fulfill this minimum requirement were judged as "0 match."

TABLE III

Location of the selected sequence elements in the genes obtained by computer surveys

	Sequences analyzed[a]	In cds[b]	In intron[b]	Downstream of cds[b]	
				Total	First[c]
Pool 1	322	104 (32)	74 (23)	144 (45)	107 (33)
Pool 2a	179	23 (13)	30 (17)	136 (70)	108 (60)
Pool 2b	61	5 (8)	26 (43)	20 (49)	14 (39)
Pool M	68	8 (12)	21 (31)	39 (57)	30 (44)

[a] 850 sequences of pool 1 with at least 10 matches, 582 sequences of pool 2a with at least 14 matches, 166 sequences of both pools 2b (T_4 and T_3) with at least 15 matches, and 175 sequences of pool M with 8 matches to the requested sequence elements were analyzed. Those sequences that were present several times, did not contain any coding sequence, or belonged to non-vertebrate viruses were eliminated. The number of the residual sequences was identical to the number of sequences examined, except for pool 1 (462 residual sequences).

[b] cds are coding sequence; numbers in brackets are percentages.

[c] Number of those homologies that belonged to the first AATAAA signal downstream of the coding sequence are shown.

their polypeptide composition differed with respect to the 64-kDa subunit (Fig. 1) which was shown to interact with the downstream elements of pre-mRNAs (31, 34). The 64-kDa polypeptide split into a 62/64-kDa doublet in HeLa CstF and a 60/62-kDa doublet in calf thymus CstF, which might be due to partial degradation. In addition, two polypeptides of 70 and 52 kDa were present in calf thymus CstF, which were recognized by anti-64-kDa polyclonal antibodies and could be UV cross-linked to RNA (Fig. 1, *B–D*). Whereas the 52-kDa protein might be a degradation product, the 70-kDa protein is an alternative form of the 64-kDa polypeptide that may result from alternative splicing. It is unlikely that its significantly different migration behavior on SDS-polyacrylamide gels is caused by post-translational modifications. Furthermore, the monoclonal antibody 3A7 (30) directed against the human 64-kDa subunit did not recognize the 70-kDa polypeptide, indicating the absence of the required epitope (data not shown). The precise nature of the new subunit will have to be determined by cDNA cloning.

It can be speculated whether alternative 64-kDa subunits might confer different RNA-binding properties to CstF. In fact, the RNA pool selected by calf thymus CstF differed from the selected HeLa pools: the uracil content was significantly higher (9.8) than in the HeLa pools (Table I, 6.9 and 7.1) and element 1 sequences were rare (pool A, 9.6%; pool B, 34%; and pool C, 41%; Fig. 2). However, no significant differences between these CstFs were observed in RNA-binding reactions or in reconstituted *in vitro* cleavage reactions with several RNA substrates (data not shown). Since both HeLa and calf thymus CstF contain a mixed population of different 64-kDa polypeptides, only separate analysis of these subunits can address this question in detail.

CstF Selects Highly Conserved RNA Ligands—In contrast to the high sequence variability of downstream elements *in vivo*, only three specific sequence elements were selected by CstF *in vitro*, element 1 (AUGCGUUCCUCGUCC) and two related elements 2a (UGUGUYN$_{0-4}$UUYAYUGYGU) and 2b (UUGYUN$_{0-4}$AUUUACU(U/G)N$_{0-2}$YCU). All selected RNAs contain either of these motifs, and only a few sequences share homologies with lower than 75% to the consensus elements 1 and 2a. Several nucleotides of these elements are highly conserved (at least 76% identity), which is surprising regarding the difficulty to determine a consensus sequence by alignments of naturally occurring downstream elements. Only element 2b is slightly less conserved. Since all purifications and selection experiments were performed independently, the enrichment of identical sequence elements in the different pools imply sequence-specific RNA-binding preferences for CstF-RNA interactions.

The selected elements share homologies to motifs that have been proposed for downstream element function. Element 1 is a significantly extended version with one mismatch of the previously proposed consensus sequence for downstream elements YGUGUUYY (8). Both elements 2a and 2b contain novel combinations of a GU-rich motif similar to GUGUUG (9) and an AY motif similar to CAYUG (10). All elements have a GU-rich motif in their 5′-half in common. These GU motifs vary slightly as follows: UGCGUU for element 1, YGUGUY for element 2a, and UUGYU for element 2b. Interestingly, point

Pool 1

element 1	A	T	G	C	G	T	T	C	C	T	C	G	T	C	C	identity (%)	distance cds...AATAAA...DE
HS299531	C	T	T	C	G	T	T	C	C	T	C	C	T	C	C	80	267 25
MM23462	T	T	G	T	G	T	T	C	C	T	C	C	T	C	C	80	270 39
HS23KDHBP	A	T	G	C	G	T	T	G	C	C	T	G	C	C	C	73	5 15
HSLUT6	A	T	G	C	A	G	T	C	C	T	C	A	T	C	T	73	91 16
HSDAFA	A	T	G	C	T	T	T	C	A	T	T	G	T	C	T	73	81 23
OCHBZ2	C	C	G	C	G	T	T	C	C	T	T	G	T	G	C	73	76 24
HSMHANTL	A	T	G	T	G	T	T	T	C	T	T	G	T	G	C	73	388 28
HSKER65D	A	T	T	C	G	C	T	C	C	A	T	G	T	C	C	73	333 31
HEVZVXX	A	T	G	C	G	G	T	T	C	C	C	T	T	C	C	73	651 34
HSHSDI09	A	T	T	C	T	T	T	C	C	T	A	C	T	C	C	73	170 37
HH43400	G	T	G	C	G	T	T	A	C	A	G	G	T	C	C	73	1153 41
HEHSSGX	G	T	G	C	G	A	T	C	C	A	C	G	C	C	C	73	59 45

Pool 2a

element 2a	Y	G	T	G	T	Y	N_{0-5}	T	T	Y	A	Y	T	G	N_{0-2}	Y	G	T	identity (%)	distance cds...AATAAA...DE
HEVZIRLS	C	G	T	G	T	C	T	T	T	T	T	T	T	G	A	T	G	T	94	9 15
AD40E1AB	T	G	T	G	T	T	T A	T	T	T	C	T	T	G	G G	C	G	T	94	10 17
AD4E1A	T	G	T	G	T	T	T T A	T	T	T	C	T	T	G	G G	T	G	T	94	12 17
HS3331710	T	G	T	G	T	T	T G T G C	T	T	T	T	C	T	G	G	T	G	T	94	103 20
HEVZVXX	T	G	T	G	T	T	T T C T	T	T	T	T	T	T	G		T	G	T	94	47 23
OCBGLO	T	G	T	G	T	T	G G A A	T	T	T	T	T	T	G		T	G	T	94	71 24
HEHS1ATI	C	G	T	G	T	T	C T T	T	T	T	A	T	C	G	C A	C	G	T	94	65 27
HSUGT03	A	G	T	G	T	T		T	T	C	A	C	T	G	G	T	G	T	94	919 27
HSIFNAA	T	G	T	G	T	T	G	T	T	C	A	T	T	G	A A	C	T	T	94	306 30
HSLDLR18	C	G	T	G	T	T	A C T G T	T	G	C	A	C	T	G	A	T	G	T	94	1907 50
BHV130KB	T	T	T	G	T	T	T G C C G	T	T	T	A	T	T	T		T	G	T	88	55 29
HSLHDC	T	G	T	G	T	C	A C	T	T	A	A	T	T	G	G C	T	G	C	88	288 30

Pool 2b

element 2b	T	T	G	Y	T	N_{0-5}	A	T	T	T	(T)	A	C	T	T/G	N_{0-2}	Y	C	T	identity (%)	distance cds...AATAAA...DE
HS0MGP	T	T	G	C	T	G C T	A	T	T	T	T	A	C	T	G		T	C	T	100	158 23
XLCRYB	T	T	G	T	T	T	A	T	T	T		A	C	T	G	T	T	C	T	100	0 36
S78854	T	T	G	C	T	T T T C	C	T	T	T	T	A	C	T	T		C	C	T	94	174 21
OAKERC2G	T	T	G	T	T	T	A	T	T	T	T	A	T	T	G	T G	C	C	T	94	352 32
EAFGP	T	T	G	C	T		A	T	T	C		A	C	T	T	A	T	C	T	94	241 5
HSATPRMR	T	A	G	T	T	A A A C	A	T	T	T		A	C	T	T		T	C	T	94	218 11
S41209	T	T	G	T	T	T G G	A	T	T	T		C	C	T	T		C	C	T	94	992 16
MH68TKH	T	T	G	T	T		T	C	T	T	T	A	C	T	G	T A	T	C	T	94	2 33
HS08023	T	T	A	T	T	T C	A	T	T	T	C	A	C	T	T	A	T	C	T	88	225 7
HSTFPB	T	T	G	T	T	A C T	G	T	T	G	T	A	C	T	T	A T	T	C	T	88	1119 12
HSNF1AA1	T	T	G	T	T	T G G T	A	T	A	T	T	A	C	T	T		T	T	T	88	136 44
ECJKCK	T	T	G	T	T	A T	A	T	A	T	T	G	C	T	T	G	T	C	T	88	180 46

FIG. 4. **Putative downstream elements are obtained by computer surveys with the individual sequences selected by CstF.** The elements used for the EMBL data library screens are given on *top* of every alignment. The percentage of identity to the sequence used for the search is indicated for every gene, and identical nucleotides are *shaded* in *gray*. The distance (nt) between the end of the coding sequence (*cds*) and the polyadenylation signal AATAAA (*distance cds . . . AATAAA*) as well as the distance between the AATAAA signal and the homologies (*distance AATAAA . . . DE*) are indicated.

mutagenesis of the downstream element of SV40 early pre-mRNA (23) that changed the natural sequence UUGUGGU to either UUGUGUU or UUGUUGU and thus created sequences identical to the selected GU motifs of element 2a and 2b, respectively (underlined), increased the 3′-end processing efficiency about 3-fold in comparison to wild type. These results and those obtained by our *in vitro* selection experiments indicate that GU motifs play a critical role in CstF-RNA interaction and that specific rather than random GU-rich sequences seem to be preferred.

In contrast to the conserved 5′ parts of all selected elements, the 3′ parts are more variable. In element 1, a pyrimidine-rich sequence is present, whereas elements 2a and 2b contain AY motifs. These findings suggest a bipartite structure and a variable sequence requirement in the 3′ part of the RNA ligands. One can speculate that the 64-kDa polypeptide of CstF binds to the RNA with two different domains since it does not only contain a ribonucleoprotein-like RNA binding domain (RBD) but also 17 RGG-like motifs preceding and overlapping with the MEAR(A/G) repeats, which have been suggested to form an α-helical structure and to be involved in protein-protein interactions (31). RGG-like motifs usually occur in proteins that also contain RBDs and are often modified post-translationally to modulate RNA-binding activity (for review, see Ref. 46). Modifications of these RGG-like motifs may result in different sequence preferences for the 3′ part of the RNA ligand. The existence of a second RNA-binding region in the 64-kDa subunit of CstF is also consistent with the results of a recent

SELEX study with the isolated RBD of the 64-kDa subunit of human CstF (47). In contrast to the sequences selected with the complete CstF factor described here, the RBD alone predominantly selected short G/U-containing sequence elements. This difference is likely due to the fact that amino acids outside of the RBD of the 64-kDa polypeptide contribute to the binding specificity of CstF.

The adenosine residues of the AY motifs present in elements 2a and 2b are highly conserved (at least 84%) and thus might be critical for CstF-RNA binding. Further evidence for an involvement of adenosine residues in CstF-RNA interaction comes from modification interference assays with the selected RNAs A-1 and A-2 (data not shown) as well as from two point mutagenesis experiments on downstream elements. It was demonstrated that a stretch of five uracil residues is sufficient to restore cleavage activity of a pre-mRNA that is otherwise not processed. Inserting adenosine residues at four of these five positions significantly decreased cleavage activity. Only the sequence UUAUU, which resembles the central AY motifs of the selected elements 2a (UYAYU) and 2b (UUACU), was processed as efficiently as UUUUU (14). Point mutagenesis of the downstream element of adenovirus E2A revealed a 1.3-fold stimulation in 3′-end processing, when the sequence UUGUUU was changed to UUAUUU (23). Since this effect was not as dramatic as changes in the GU-rich motif of SV40 early pre-mRNA, GU-rich motifs obviously play a more critical role in 3′-end processing than AY motifs. This is also indicated by the finding that all selected elements contained a GU-rich element but not all had an AY-rich motif.

The Selected Elements Function as Downstream Elements in 3′-End Processing—To investigate whether the selected sequences were able to function as downstream elements in 3′-end processing, they were subcloned into an SV40 late pre-mRNA derivative whose polyadenylation signals had been inactivated by deleting the natural downstream region. All selected sequences tested, including shortened versions, were able to restore cleavage activity, although to different extents. Those RNAs (SV-A42 and SV-E1P12), whose 11-mer element was located as far downstream from the AAUAAA signal as the CstF-binding site of SV40 (34), were processed more efficiently than substrates that contained the 11-mer element further upstream (SV-B13, SV-C1, and SV-E1P4). This is in agreement with previous reports that showed the dependence of both efficiency and accuracy of the cleavage reaction on the position of the downstream element (14, 19, 21, 22, 24, 25). Nevertheless, the only RNA that was cleaved with wild type efficiency and accuracy was SV-E1d, which contained a duplication of the 11-mer element and thus created a bipartite downstream element. Bipartite downstream elements have not only been reported for SV40 late RNA (17) but also for other RNAs (19, 20) and support the idea that CstF contains two RNA-binding domains.

Selected element 2a sequences were also able to restore cleavage activity. The efficiencies of these RNA substrates were comparable to those element 1 constructs of which the 11-mer element was located at the beginning of the inserted sequence. This is most probably due to a non-optimal position of the downstream element relative to the AAUAAA signal.

Further analysis revealed that even short GU and AY motifs were able to restore cleavage activity. Again, the GU motif was the most important part of element 2a, since it could substitute for the AY motif but not vice versa. This is in good agreement with the conservation of the GU motif in the 5′ part of all selected elements and the already suggested role for GU-rich sequences in downstream element function (see above). Furthermore, our results demonstrate that CstF-RNA interactions during 3′-end processing tolerate significant mutations of the downstream element. This is in contrast to the highly conserved sequences of the elements that were selected by CstF *in vitro* in the absence of any other 3′-end processing factor. It is likely that protein-protein interactions between CstF and other components of the 3′-end processing machinery can compensate for weak CstF-RNA interactions. Therefore, several sequences can function as downstream elements although with different efficiencies. This might enable the cell to carefully regulate 3′-end processing. It has been demonstrated that overexpression of the 64-kDa subunit in stably transformed B-cells induced alternative polyadenylation (37). Considering the different polypeptide compositions of calf thymus and HeLa CstF, it is also conceivable that this regulation might be influenced by the expression of different 64-kDa subunits.

Sequence Homologies to the Selected Elements Are Present in Many Genes—A computer survey of the EMBL data library was performed to investigate whether the sequence elements selected by CstF *in vitro* were also present in genes and thus play a role in 3′-end processing *in vivo*. Homologies to either of the selected elements 1, 2a, or 2b are present in 89% of all sequences with a perfect match to the AATAAA hexamer (pool V). Taking into account that only 16% of all AATAAA signals are present in coding sequences (48), about 70% of all homologies found should be located outside of protein coding sequences. Indeed, element 2a was mainly found in the 3′-UTR of genes. This strongly suggests a role for element 2a in 3′-end processing *in vivo*, particularly if one takes into account that 3′-UTRs are four to five times less abundant in this sequence library than coding sequences. In contrast, element 1 was also frequently found within protein coding sequences, a finding that does not strongly argue for its involvement as a general downstream element *in vivo* on first sight. But since these homologies included the AATAAA hexamer, the presence of these sequences within the coding region does not exclude the function of such a sequence in 3′-end processing when appropriately located in 3′-UTRs.

Furthermore, 128 downstream element regions (8, 34) were screened for the presence of the selected elements 2a and 2b (data not shown). About 51% of these sequences contained the selected elements with at least 70% identity downstream of their natural cleavage sites. Two of these sequences were 94% identical to element 2a and were also detected with the computer screen (Fig. 4, *OCBGLO* and *HEHS1ATI*). Interestingly, a detailed study of the rabbit β-globin pre-mRNA (Fig. 4, *OCBGLO*) identified the sequence that exhibits the high homology to element 2a as the natural downstream element (19) .

Conclusions—Our results demonstrate that CstF purified from two sources differed with respect to their 64-kDa subunit that is responsible for CstF-RNA interactions. Calf thymus CstF contained an additional, novel 70-kDa polypeptide that could be UV cross-linked to RNA and that was recognized by polyclonal antibodies directed against the 64-kDa subunit. Considering the sequence variability of downstream elements, the selection of highly conserved sequence elements by CstF was surprising. The selected motifs functioned as downstream elements in *in vitro* 3′-end processing reactions. Homologies to all selected elements were found in the 3′-UTRs of many genes. These results strongly suggest that the sequences selected *in vitro* function as natural downstream elements *in vivo*. We propose that the closely related elements 2a and 2b represent a novel consensus sequence for downstream elements.

Acknowledgments—We thank Marvin Wickens for the plasmid pSV-141/-1; Georges Martin for recombinant bovine PAP; Elmar Wahle, Andreas Jenny, and Silvia Barabino for purified CPSF; Christine Milcarek and Kathleen Martincic for the polyclonal antibodies directed against the 64-kDa subunit of CstF; Clinton MacDonald for the mono-

clonal antibody directed against the 64-kDa subunit of CstF and sharing sequence information on downstream elements. We also thank Silvia Barabino, Lionel Minvielle-Sebastia, Mary O'Connell, Ursula Rüegsegger, and Elmar Wahle for comments on the manuscript. T. D. thanks Iain Mattaj and Gene Expression (EMBL) and Heiner Schirmer for stimulating discussions.

REFERENCES

1. Keller, W. (1995) *Cell* **81,** 829–832
2. Keller, W., and Minvielle-Sebastia, L. (1997) *Curr. Opin. Cell Biol.* **9,** 329–336
3. Wahle, E., and Keller, W. (1996) *Trends Biochem. Sci.* **21,** 247–250
4. Wahle, E. (1995) *Biochim. Biophys. Acta* **1261,** 183–194
5. Wahle, E., and Keller, W. (1994) in *RNA Processing: A Practical Approach* (Hames, B. D., and Higgins, S. J., eds) pp. 1–34, Oxford University Press, New York
6. Manley, J. L. (1995) *Curr. Opion. Genet. & Dev.* **5,** 222–228
7. Wahle, E., and Keller, W. (1992) *Annu. Rev. Biochem.* **61,** 419–440
8. McLaughlan, J., Gaffney, D., Whitton, J. L., and Clements, J. B. (1985) *Nucleic Acids Res.* **13,** 1347–1368
9. Taya, Y., Devos, R., Travernier, J., Cheroutre, H., Engler, G., and Fiers, W. (1982) *EMBO J.* **1,** 953–958
10. Berget, S. (1984) *Nature* **309,** 179–182
11. Sadofsky, M., Connelly, S., Manley, J. L., and Alwine, J. C. (1985) *Mol. Cell. Biol.* **5,** 2713–1719
12. Sittler, A., Gallinaro, H., and Jacob, M. (1994) *Nucleic Acids Res.* **22,** 222–231
13. Birnstiel, M. L., Busslinger, M., and Strub, K. (1985) *Cell* **41,** 349–359
14. Chou, Z. F., Chen, F., and Wilusz, J. (1994) *Nucleic Acids Res.* **22,** 2525–2531
15. Conway, L., and Wickens, M. (1987) *EMBO J.* **6,** 4177–4184
16. Gimmi, E. R., Soprano, K. J., Rosenberg, M., and Reff, M. E. (1988) *Nucleic Acids Res.* **16,** 8977–8997
17. Zarkower, D., and Wickens, M. (1988) *J. Biol. Chem.* **263,** 5780–5788
18. Ryner, L. C., Takagaki, Y., and Manley, J. L. (1989) *Mol. Cell. Biol.* **9,** 1759–1771
19. Gil, A., and Proudfoot, N. J. (1987) *Cell* **49,** 399–406
20. Chen, J. S., and Nordstrom, J. L. (1992) *Nucleic Acids Res.* **20,** 2565–2572
21. Woychick, R. P., Lyons, H. H., Post, L., and Rottman, F. M. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81,** 3944–3948
22. Zhang, F., Denome, R. M., and Cole, C. N. (1986) *Mol. Cell. Biol.* **6,** 4611–4623
23. McDevitt, M. A., Hart, R. P., Wong, W. W., and Nevins, J. R. (1986) *EMBO J.* **5,** 2907–2913
24. Mason, P. J., Elkington, J. A., Lloyd, M. M., Jones, M. B., and Williams, J. G. (1986) *Cell* **46,** 263–270
25. Goodwin, E. C., and Rottman, F. M. (1992) *J. Biol. Chem.* **267,** 16330–16334
26. Manley, J. L., and Takagaki, Y. (1996) *Science* **274,** 1481–1482
27. Gilmartin, G. M., and Nevins, J. R. (1989) *Genes Dev.* **3,** 2180–2189
28. Gilmartin, G. M., and Nevins, J. R. (1991) *Mol. Cell. Biol.* **11,** 2432–2438
29. Weiss, E. A., Gilmartin, G. M., and Nevins, J. R. (1991) *EMBO J.* **10,** 215–219
30. Takagaki, Y., Manley, J. L., MacDonald, C. C., Wilusz, J., and Shenk, T. (1990) *Genes Dev.* **4,** 2112–2120
31. Takagaki, Y., MacDonald, C. C., Shenk, T., and Manley, J. L. (1992) *Proc. Natl. Acad. Sci. U. S. A.* **89,** 1403–1407
32. Takagaki, Y., and Manley, J. L. (1992) *J. Biol. Chem.* **267,** 23471–23474
33. Takagaki, Y., and Manley, J. L. (1994) *Nature* **372,** 471–474
34. MacDonald, C. C., Wilusz, J., and Shenk, T. (1994) *Mol. Cell. Biol.* **14,** 6647–6654
35. Mann, K. P., Weiss, E. A., and Nevins, J. R. (1993) *Mol. Cell. Biol.* **13,** 2411–2419
36. Edwalds-Gilbert, G., and Milcarek, C. (1995) *Mol. Cell. Biol.* **15,** 6420–6429
37. Takagaki, Y., Seipelt, R. L., Peterson, M. L., and Manley, J. L. (1996) *Cell* **87,** 941–952
38. Tuerk, C., and Gold, L. (1990) *Science* **249,** 505–510
39. Bienroth, S., Wahle, E., Suter-Crazzolara, C., and Keller, W. (1991) *J. Biol. Chem.* **266,** 19768–19776
40. Martin, G., and Keller, W. (1996) *EMBO J.* **15,** 2593–2603
41. Tuerk, C., Eddy, S., Parma, D., and Gold, L. (1990) *J. Mol. Biol.* **213,** 749–761
42. Rüegsegger, U., Beyer, K., and Keller, W. (1996) *J. Biol. Chem.* **271,** 6107–6113
43. Dandekar, T., and Hentze, M. W. (1995) *Trends Genet.* **11,** 45–50
44. Kahn, P., and Cameron, G. (1990) *Methods Enzymol.* **183,** 23–31
45. Takagaki, Y., Ryner, L. C., and Manley, J. L. (1989) *Genes Dev.* **3,** 1711–1724
46. Burd, C. G., and Dreyfuss, G. (1994) *Science* **265,** 615–621
47. Takagaki, Y., and Manley, J. L. (1997) *Mol. Cell. Biol.* **17,** 3907–3914
48. Day, I. N. M. (1992) *Gene* (*Amst.*) **110,** 245–249