Nucleic Acids Research, 1995, Vol. 23, No. 14 2629–2635

Cloning of cDNAs encoding the 160 kDa subunit of the bovine cleavage and polyadenylation specificity factor

Andreas Jenny and Walter Keller*

Department of Cell Biology, Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

Received April 20, 1995; Revised and Accepted June 5, 1995

EMBL accession no. X83097

ABSTRACT

3′-processing of mRNA precursors depends on several protein factors. One of them, cleavage and polyadenylation specificity factor (CPSF) is required for the cleavage of the mRNA precursor as well as for the tail elongation reaction. We have obtained complementary DNA encoding the 160 kDa subunit, which had previously been shown to interact with the AAUAAA polyadenylation signal. The cDNAs code for an open reading frame of 1444 amino acids. The translated protein has a calculated molecular weight of 161 kDa and a predicted pI of 6.2. Polyclonal antibodies raised against a bacterially expressed fragment of the cDNA recognise the 160 kDa subunit of purified calf thymus CPSF. The sequence contains a possible nuclear localisation signal but none of the known RNA binding motifs. It does, however, show sequence similarities to a UV-damaged DNA binding protein (UVdDb).

INTRODUCTION

Most 3′ ends of eukaryotic messenger RNAs are generated by endonucleolytic cleavage and subsequent addition of a poly(A) tail of ~250 residues [for reviews see (1–3)]. The poly(A) tail is thought to be important for the turnover of the mRNA, for its export from the nucleus and for translation (3–5). *In vivo*, cleavage and polyadenylation are tightly coupled and depend on protein factors as well as signals on the RNA transcript. In mammals, the processing reaction depends on the canonical AAUAAA polyadenylation signal that is located 10–35 nucleotides upstream of the cleavage site, and on a less well defined U or G/U rich downstream element. The AAUAAA sequence is highly conserved and very responsive to mutations and modifications (6,7). *In vitro*, the polyadenylation reaction can be uncoupled from the cleavage reaction and both steps can therefore be studied separately (8). The cleavage of the precursor depends on the cleavage and polyadenylation specificity factor [CPSF; (9–12)], the cleavage stimulation factor [CstF; (13)], cleavage factors I and II [CFI/II; (12,14)] and poly(A) polymerase [PAP; (12,14)]. The second reaction step, processive addition of the poly(A) tail, requires CPSF, PAP and poly(A) binding protein II [PAB II; (15,16)]. Among the cleavage factors, direct interaction with the RNA substrate has been shown for CstF and CPSF (13,17–19). CstF binds to the downstream element, as indicated by UV-crosslinking and RNAse H digestion (20), whereas the AAUAAA signal is directly bound by CPSF (17,21).

CPSF has been purified from calf thymus and HeLa cells and consists of four polypeptides with apparent molecular masses of 160, 100, 73 and 30 kDa, only three of which are probably essential for the activity *in vitro* (9,11). So far, only the 100 kDa subunit has been cloned (21). UV-crosslinking experiments with RNA substrates containing a functional polyadenylation signal have shown that the 160 kDa subunit is in close contact with the RNA (17,18,21). It is therefore very likely that this subunit is involved in the RNA recognition step of the polyadenylation reaction.

Here we report the cloning of cDNAs coding for the 160 kDa subunit of CPSF. Its amino acid sequence shows similarities to a UV-damaged DNA binding protein (UVdDb) from the African green monkey. The C-terminal region can be aligned to several additional proteins, some of which have already been proposed to share similarities with UVdDb. Surprisingly, the sequence does not contain a known RNA binding motif. Polyclonal antibodies were raised against a bacterially expressed fragment of the cDNA. These antibodies detect the 160 kDa subunit of CPSF in partially purified fractions as well as in samples immunoprecipitated with antibodies raised against the 100 kDa subunit of CPSF.

METHODS

Oligonucleotides

L385′: CC*GAATTC*AARATHGGIACIACICC (6-fold degenerated);
L383′P1: CC*GTCGAC*RAARTGIGCIGTIACICG (4-fold degenerated);
L383′P2: CC*GTCGAC*RAARTGIGCIGTIACYCT (8-fold degenerated);
Anchor: TTA*GCGGCCGC*CTTTAGTGAGGGTTAATTTCGTddA;
T3LZAP: CGAAATTAACCCTCACTAAAG;
N1Race: CGAGGCCGGTGGGCGGATGC;
NS1: CCTGCTTGTACACAGCGTAC.
R = A, G; Y = T, C; H = A, T, C; I = inosine. Letters in italics correspond to added restriction sites.

* To whom correspondence should be addressed

Protein sequencing

CPSF was purified from calf thymus total extract as described previously (9). CPSF [poly(A) Sepharose pool] corresponding to 54 μg of the 160 kDa subunit (340 pmol) was concentrated in Centricon-30 microconcentrators (Amicon), separated on a preparative 6% SDS–polyacrylamide gel (22) and blotted onto nitrocellulose with a semi-dry blotter (23). The blot was stained with Ponceau S (Sigma), the band migrating at 160 kDa excised (30 mm × 3.5 mm) and washed with H_2O. *In situ* protein digestion with endo-LysC, HPLC purification and peptide sequencing was performed by Dr William Lane, Harvard Microsequencing Facility, Cambridge, MA.

Reverse transcription and polymerase chain reaction

Both ends of peptide L38 (23 amino acids) were used to design PCR primers containing inosines at positions of 4-fold degeneracy. One microgram of calf thymus poly(A)$^+$ RNA (24) was reverse transcribed with either random hexamer primers (40 ng), oligo-dT (500 ng), oligonucleotide L383′P1 or L383′P2 (70 ng each) and Superscript reverse transcriptase (Gibco), as recommended by the manufacturer. The cDNA was purified with the Qiaquick spin PCR product purification kit (Diagen). Five percent of the cDNA and 50 pmol of each primer were used per 50 μl PCR reaction. Amplification of the cDNA was done in an OmniGene thermal cycler (AMS) with addition of Supertaq polymerase (Stehelin, Basel) after an initial denaturation step of 3 min at 94°C. Denaturation was for 30 s at 94°C, extension for 30 s at 72°C. The annealing condition of the first 5 cycles was 38°C for 30 s, for the 30 following cycles annealing was done at 50°C for 30 s. After the last cycle, the reaction was incubated for 5 min at 72°C. The reaction products were precipitated with ethanol and loaded on a 8% polyacrylamide gel in 0.5× TBE (24). After staining the gel with ethidium bromide, the band with the expected size of 85 nt was excised, reamplified and gel-purified again. The PCR product was sequenced directly with 50 ng of the PCR primers L385′ and L383′P1.

Isolation of cDNAs encoding the 160 kDa subunit of CPSF and Northern blots

5×10^5 colonies of an oligo-dT primed calf thymus cDNA plasmid library (21) and 7.5×10^5 p.f.u. of a random and oligo-dT primed calf aorta endothelial cDNA library in λZAPII (Stratagene # 936705) were screened with the PCR product as probe. The phage library was screened in duplicate at a density of 5×10^4 p.f.u. per 150 mm plate with Hybond N$^+$ filters (Amersham). The plasmid library was screened at a density of 2.5×10^4 colonies per 150 mm plate with reinforced nitrocellulose filters (Schleicher and Schuell). Filter lifting procedures were as suggested by the manufacturers, except that the colony lifts were neutralised twice with 1 M ammonium acetate (pH 7.5), followed by a treatment with $CHCl_3$ for 45 s (24) and baking in a vacuum oven at 80°C for 2 h.

After denaturation, 50 ng PCR product was labeled with Klenow enzyme with L385′ and L383′P1 as primers and [α-^{32}P]dATP for 30 min at 30°C. The reaction was then heat-denatured again and fresh enzyme was added. After another 30 min at 37°C, the reaction was chased with 1 μl of 10 mM dNTPs for 10 min and purified over a Sephadex G-25 spin column. The filters were hybridised overnight at 42°C in 40% formamide, 5× Denhardt's, 6× SSC, 1% SDS, 1 mM EDTA, 100 μg/ml denatured salmon sperm DNA (24). They were then washed twice for 15 min at room temperature in 2× SSC, 0.1% SDS followed by 20 min at 60°C in 1× SSC, 0.1% SDS and exposed overnight to Kodak X-OMAT AR films with an intensifying screen. Five clones from the calf thymus library were purified and analysed. The originally 39 positive clones of the aorta library were characterised by PCR (25) and the longest ones purified and mapped. ExoIII deletions (26) of the longest clone, CA#15, were prepared (Erase-a-Base system, Promega) and sequenced.

The N-terminal *Eco*RI (linker) *Sac*I fragment of CA#15 (Fig. 1) was used as a probe to rescreen the calf aorta library under high stringency conditions [50% formamide, (24)]. The resulting 22 clones were again characterised by PCR and the longest ones were plaque purified. The clone extending farthest into the N-terminal region (N#1) was sequenced. To obtain sequence lying more 5′ to N#1, ligation-anchored RACE was performed as described (27), except that the cDNA was purified over a PCR Qiaquick spin column (Diagen) before ligation to the anchor. After the ligation, the cDNA was purified twice over a PCR Qiaquick spin column in the presence of 10 μg tRNA to remove any residual oligonucleotide. Two rounds of amplification (35 cycles each) were performed at an annealing temperature of 50°C, each with the primer T3LZAP (complementary to the anchor) and the nested primers N1RACE (nucleotides 174–155) and NS1 (nucleotides 153–134), respectively. The products were gel purified, reamplified and subcloned into the *Not*I and *Eco*RV sites of pBluescript-IISK$^+$ (Stratagene) and sequenced.

All sequences were determined on an Applied Biosystems 373A sequencer with dye terminators according to the manufacturers instructions and assembled with the SeqMan program (DNA-Star). Every base was sequenced at least once on both strands. DNA and translated protein sequences were analysed with the GCG software package (28). Databases were searched with FASTA and TFASTA (29).

For Northern blot analysis, 1 μg calf thymus poly(A)$^+$ RNA was separated on a denaturing formaldehyde gel (24), blotted onto Hybond N$^+$ membrane (Amersham) and UV crosslinked in a Stratalinker (Stratagene). Digoxygenine-labelled antisense RNAs were transcribed from *Kpn*I (polylinker) linearised clones CA#15 and N#1 with T3 RNA polymerase. Hybridisation and chemiluminescent detection were performed according to the instructions of the manufacturer (Boehringer Mannheim), except that the blocking time before the antibody incubation was extended to 3 h. The blots were exposed to Kodak XOMAT AR film for 15 min.

Production of polyclonal antibodies, immunoprecipitation and Western blotting

The C-terminal *Bam*HI (Fig. 1) to *Hin*dIII (polylinker) of clone CA#15 was cloned in frame into the corresponding sites of pQE11 (Diagen) behind a tag of six histidines. The construct was transformed into *Escherichia coli* strain SG13009 (Diagen) containing the lac repressor plasmid pREP4. LB cultures (500 ml) were induced at an OD_{600} of 0.7 with 1 mM isopropyl-β-D-galactopyranoside and grown for additional 5 h at 37°C. The fusion protein was purified under denaturing conditions over a Ni(II)-nitrilotriacetic acid agarose column as suggested by the manufacturer (Diagen). A New Zealand white rabbit was immunised with 100 μg fusion protein emulsified with Freund's complete adjuvant. After 4 weeks, a booster injection of 20 μg protein (SDS–polyacrylamide gel purified), emulsified with 600 μl

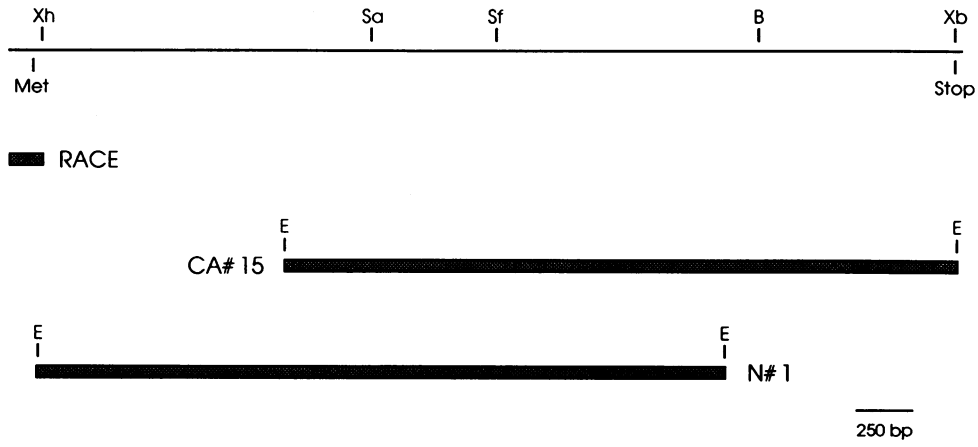

Figure 1. Alignment of the cDNA clones used for sequencing and generation of a full-length clone. Restriction sites are indicated on top, except for the cloning sites of each clone. Met indicates the start codon, stop the stop codon. B, *Bam*HI; E, *Eco*RI; Sa, *Sac*I; Sf, *Sfi*I; Xb, *Xba*I; Xh, *Xho*I.

Specol (Central Veterinary Institute, Netherlands) was given. In intervals of 4 weeks, two additional booster injections of 100 µg fusion protein emulsified with Specol were applied. The serum (serum # 0680) was affinity-purified on antigen adsorbed to nitrocellulose (30).

For Western blotting, CPSF samples were separated over a 9% SDS–polyacrylamide gel and transferred to nitrocellulose with a semi-dry system (23). The blots were blocked with 5% w/v low fat dry milk in phosphate buffered saline (PBS), 0.05% v/v Tween 20 (PBST). They were then incubated in a 1:500 dilution of the affinity purified antibody in PBST for 3 h, washed four times for 15 min with PBST, incubated with horseradish peroxidase-coupled swine anti-rabbit antibody (DAKO) diluted 1:2000 in PBST for 1 h and washed again with four changes of PBST. The signal was detected with the ECL system (Amersham).

Immunoprecipitation of purified CPSF [Superose 6 fraction; (9,21)] was done as described in Jenny *et al.* (21) with an equivalent of 200 ng of the 160 kDa subunit.

In vitro translation

A construct containing the entire open reading frame of the 160 kDa subunit of CPSF was made by replacing the *Sfi*I (Fig. 1) *Not*I (polylinker) fragment of clone N#1 with the *Sfi*I *Not*I fragment of CA#15. The construct was linearised with *Not*I and transcribed with T7 RNA polymerase. The RNA was translated in rabbit reticulocyte lysate and labeled with [^{35}S]methionine according to the instructions of the manufacturer (Promega). After separation, SDS–polyacrylamide gels were fixed, incubated in 1 M Na-salicylate, dried and exposed to Kodak XOMAT-AR films.

Polyadenylation assays

A Blue Sepharose pool of CPSF (9) obtained from 2 kg calf thymus was loaded onto a 250 ml Heparin Sepharose column and eluted with a six column volume gradient of 50–300 mM KCl (9). Aliquots (2 µl) of the 20 ml fractions were loaded on a 9% SDS–polyacrylamide gel and used for Western blot analysis. Aliquots (0.5 µl) of the fractions were assayed for specific polyadenylation activity as described earlier (16), except that the incubation temperature was 30°C.

RESULTS AND DISCUSSION

Cloning and sequencing of the 160 kDa subunit of CPSF

The cleavage and polyadenylation specificity factor (CPSF) was purified as described previously (9). Approximately 340 pmol CPSF corresponding to 54 µg 160 kDa subunit were separated on a preparative 6% SDS–polyacrylamide gel and blotted onto nitrocellulose. The protein was digested *in situ* with lysyl-endopeptidase, and the resulting peptides were purified by reverse phase HPLC chromatography. Only four peptide-containing fractions could be identified and were sequenced. Two of the fractions gave double sequences. The six sequenced peptides are indicated in Figure 3 (underlined twice). Peptide L38, which turned out to be at the C-terminus, was used to design PCR primers that were only 4- (L383'P1) to 8- (L383'P2) fold degenerate and that contained up to four inosines (L383'P1). Amplification of calf thymus cDNA with these primers gave a band with the expected length of 85 nt. After reamplification, the fragment was sequenced directly with the PCR primers to confirm that it encoded the internal amino acids of the sequenced peptide. Using this PCR fragment as probe, we screened a calf thymus and a calf aorta cDNA library. The longest clone (CA#15; Fig. 1) contained an insert of 3 kb. Its N-terminal *Sac*I fragment was used to screen the calf aorta library again in order to find cDNA clones covering the N-terminus of the CPSF 160 kDa subunit. The clone extending the most in the N-terminal direction (N#1; Fig 1) contained an insert of 3.3 kb. Both cDNA clones were sequenced completely from nested deletions.

A Northern blot analysis of calf thymus poly(A)$^+$ RNA confirmed that both clones recognised the same mRNAs and that neither of them contained unrelated gene fusions. Both probes revealed two strong bands of 5.1 and 5.7 kb and a less prominent band of 6.5 kb (Fig. 2). As judged by the short exposure time of 15 min, the transcripts are abundant. The open reading frame (see below) was not blocked in front of the potential initiation codon. Since two additional rounds of cDNA library screening with N-terminal probes did not extend the sequences at the 5′ end considerably (data not shown), we amplified cDNA extending towards the 5′ region by ligation-anchored RACE (27). The longest product we were able to amplify extended the known

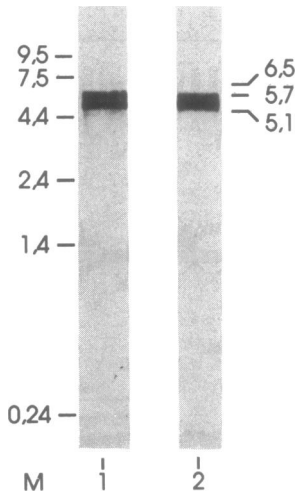

Figure 2. Northern blot analysis of clones CA#15 and N#1 (see Materials and Methods): Both overlapping clones recognise the same messages of 6.5, 5.7 and 5.1 kb. Calf thymus poly(A)$^+$ RNA was probed with antisense RNA probes of clone CA#15 (lane 1) and N#1 (lane 2). Size markers (lane M) are indicated in kb.

sequence by 126 nt (RACE in Fig. 1), but still contained no stop codon in front of the presumed initiation codon.

The assembled clones span 4488 bp and encode a protein of 1444 amino acids (Fig. 3) with a predicted molecular weight of 161 213 Da and a pI of 6.4. The cDNA clones were random-primed and contain only 22 bp 3′ untranslated region, which is probably the main reason for the discrepancy between the cDNA length and the message sizes detected on the Northern blot. The open reading frame contains all six endoLys-C peptides (underlined twice in Fig. 3). Peptide L38 turned out to be the C-terminus and accordingly, it does not end with a lysine. In addition, the reading frame contains six tryptic peptides from a previous sequencing attempt (S. Bienroth, P. Jenö and W. K., unpublished; underlined once in Fig. 3). The identified peptides are equally dispersed over the entire open reading frame, which is a good indication for the accuracy of the sequence. The amino acid composition and distribution is that of an average protein except that lysines are underrepresented [63%; (31)]. The open reading frame does not encode any of the tryptic peptide fragments obtained from the 30 kDa subunit of CPSF (not shown).

Identity of the encoded protein

To prove the identity of the cloned protein, we subcloned the C-terminal *Bam*HI fragment of clone CA#15 (Fig. 1) into a HIS-tag expression vector. The resulting protein fragment of 38 kDa was overexpressed in *Escherichia coli*, purified on a Ni-NTA column and used to generate polyclonal antibodies. Purified CPSF [Superose 6 fraction; (9)] was then immunoprecipitated with the monoclonal antibody J1/27 recognizing the 100 kDa subunit of CPSF (21). Since this antibody precipitates all four subunits of CPSF (21), the Western blot was then probed with the serum raised against the expressed fusion protein. As shown in Figure 4A, the antibodies recognise a protein of 160 kDa in purified (Superose 6; lane 1) as well as in immunoprecipitated CPSF (lane 2). No signal was detected when the immunoprecipitation was carried out with an unrelated monoclonal antibody (lane 3) or when no CPSF was added to the immunoprecipitation (lane 4). The signal detected at lower molecular weights is due to a weak cross-reactivity of the peroxidase-labeled secondary antibody with the monoclonal antibody used for the immunoprecipitation.

```
   1  MYAVYKQAHP PTGLEFSMYC NFFNNSERNL VVAGTSQLYV YRLNRDSEAP   50
  51  TKNDRSTDGK AHREHREKLE LVASFSFFGN VMSMASVQLA GAKRDALLLS  100
 101  FKDAKLSVVE YDPGTHDLKT LSLHYFEEPE LRDGFVQNVH TPRVRVDPDG  150
 151  RCAAMLIYGT RLVVLPFRRE SLAEEHEGLV GEGQRSSFLP SYIIDVRALD  200
 201  EKLLNIVDLQ FLHGYYEPTL LILFEPNQTW PGRVAVRQDT CSIVAISLNI  250
 251  TQKVHPVIWS LTSLPFDCTQ ALAVPKPIGG VVIFAVNSLL YLNQSVPPYG  300
 301  VALNSLTTGT TAFPLRTQEG VRITLDCAQA AFISYDKMVI SLKGGEIYVL  350
 351  TLITDGMRSV RAFHFDKAAA SVLTTSMVTM EPGYLFLGSR LGNSLLLKYT  400
 401  EKLQEPPAST AREAADKEEP PSKKKRVDAT TGWSGSKSVP QDEVDEIEVY  450
 451  GSEAQSGTQL ATYSFEVCDS ILNIGPCANA AMGEPAFLSE EFQNSPEPDL  500
 501  EIVVCSGYGK NGALSVLQKS IRPQVVTTFE LPGCYDMWTV IAPVRKEQEE  550
 551  TLKGEGTEPE PGAPEAEDDG RRHGFLILSR EDSTMILQTG QEIMELDASG  600
 601  FATQGPTVFA GNIGDNRYIV QVSPLGIRLL EGVNQLHFIP VDLGSPIVQC  650
 651  AVADPYVVIM SAEGHVTMFL LKNDSYGGRH HRLALHKPPL HHQSKVITLC  700
 701  VYRDVSGMFT TESRLGGVRD ELGGRGGPEA EGQGAETSPT VDDEEEMLYG  750
 751  DSGSLFSPSK EEARRSSQPP ADRDPAPFRA EPTHWCLLVR ENGAMEIYQL  800
 801  PDWRLVFLVK NFPVGQRVLV DSSFGQPTTQ GEARKEEATR QGELPLVKEV  850
 851  LLVALGSRQR RPYLLVHVDQ ELLIYEAFPH DSQLGQGNLK VRFKKVPHNI  900
 901  NFREKKPKPS KKKAEGGSTE EGTGPRGRVA RFRYFEDIYG YSGVFICGPS  950
 951  PHWLLVTGRG ALRLHPMGID GPIDSFAPFH NINCPRGFLY FNRQGELRIS 1000
1001  VLPAYLSYDA PWPVRKIPLR CTAHYVAYHV ESKVYAVATS TSTPCTRVPR 1050
1051  MTGEEKEFET IERDERYVHP QQEAFCIQLI SPVSWEAIPN ARIELEEWEH 1100
1101  VTCMKTVSLR SEETVSGLKG YVAAGTCLMQ GEEVTCRGRI LIMDVIEVVP 1150
1151  EPGQPLTKNK FKVLYEKEQK GPVTALCHCN GHLVSAIGQK IFLWSLRASE 1200
1201  LTGMAFIDTQ LYIHQMISVK NFILAADVMK SISLLRYQEE SKTLSLVSRD 1250
1251  AKPLEVYSVD FMVDNAQLGF LVSDRDRNLM VYMYLPEAKE SFGGMRLLRR 1300
1301  ADFHVGAHVN TFWRTPCRGA AEGPSKKSVV WENKHITWFA TLDGGIGLLL 1350
1351  PMQEKTYRRL LMLQNALTTM LPHHAGLNPR AFRMLHVDRR VLQNAVRNVL 1400
1401  DGELLNRYLY LSTMERGELA KKIGTTPDII LDDLLETDRV TAHF        1444
```

Figure 3. Predicted amino acid sequence of the 160 kDa subunit of CPSF. The 4332 bp open reading frame was translated into a protein sequence of 1444 amino acids. The peptide sequences underlined twice correspond to the endoLys-C fragments, the sequences underlined once to tryptic fragments. Letters in colour represent the potential nuclear localisation signal (residues 894–909).

A cDNA clone was constructed containing the complete open reading frame. RNA transcribed from this construct was then translated *in vitro* in the presence of labeled [^{35}S]methionine. The translation products were either immunoprecipitated with the monoclonal antibody J1/6 that recognizes the 160 kDa subunit of CPSF (21), or separated directly on a 6% SDS–polyacrylamide gel. Lane 4 in Figure 4B shows the complete *in vitro* translation reaction with its largest band exactly comigrating with the 160 kDa subunit of purified CPSF, which had been loaded in lane 3. The migration of the three large subunits of CPSF was marked after staining the gel with Coomassie Brilliant Blue; however, the gel was destained during the salicylation procedure. Therefore, the

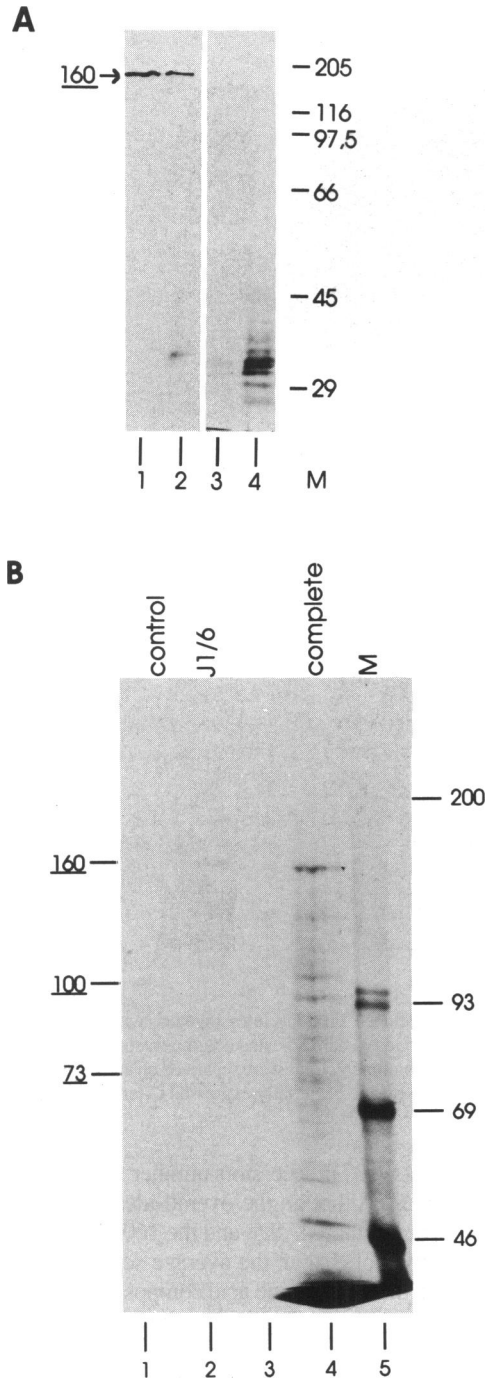

Figure 4. Immunodetection and *in vitro* translation of the CPSF 160 kDa subunit. (**A**) CPSF was immunoprecipitated with mAB J1/27 that recognises the 100 kDa subunit of CPSF and separated on a 6% SDS–polyacrylamide gel. The Western blot was probed with affinity-purified antibodies against the bacterially expressed C-terminus of the 160 kDa subunit. The 160 kDa subunit is recognised by the antiserum (lane 2). Lane 1, purified CPSF; lane 3, immunoprecipitation with an unrelated mAB (anti splicing factor $3A^{66}$ (43); lane 4, no CPSF added to immunoprecipitation. Relative molecular masses are given in kDa. (**B**) The assembled full-length clone was transcribed and translated *in vitro* and either immunoprecipitated with mAB J1/6 recognising the 160 kDa subunit of CPSF (lane 2) or loaded directly (lane 4) on a 6% SDS–polyacrylamide gel. Lane 1, immunoprecipitation with an unrelated mAB; lane 3, purified CPSF. The Coomassie Brilliant Blue stain was removed during the salicylation procedure (see text). The relative molecular masses of the standards (lane 5) are given in kDa.

bands cannot be seen on the photograph. Because many intermediate translation products were visible, we used the monoclonal antibody J1/6 to immunoprecipitate the 160 kDa protein from the translation mixture (lane 2). An unrelated monoclonal antibody did not immunoprecipitate any of the *in vitro* translated proteins (lane 1). Thus, the results shown in Figure 4 clearly demonstrate that the cloned cDNAs code for the 160 kDa subunit of CPSF.

As mentioned above, the potential translation start codon is not preceded by a stop codon upstream in the sequence. The N-terminus is blocked and could not be sequenced on the protein level (S. Bienroth, P. Jenö, A. J. and W. K., unpublished results). Nevertheless, we consider this start to be authentic for the following reasons. The known 131 bp of the leader sequence is very G/C rich (74%), which is a commonly found feature of eukaryotic genes (32). This high G/C content is probably also the reason why it was difficult to find sequences extending upstream of the initiation codon. Indeed, primer extension products were shorter than the RACE product (data not shown), probably because of spontaneous stops of the reverse transcriptase due to secondary structures formed by the G/C rich sequence. The nucleotides surrounding the methionine (AGC GCC *ATG* T) resemble the consensus proposed by Kozak (33) (GCC RCC *ATG* G). The most important nucleotide at position –3 matches this consensus. The severest deviation from the consensus is the T at position +4 that occurs with a frequency of 15% (32). Furthermore, the predicted molecular weight of 161.2 kDa agrees well with the apparent molecular mass of 160–165 kDa on SDS–polyacrylamide gels (9,11). The *in vitro* translated protein comigrates with the 160 kDa subunit of purified CPSF on 6% SDS–polyacrylamide gels, on which the resolution of high molecular weight proteins is good (Fig. 4B). We therefore assume that we have identified the correct initiation codon.

We have also investigated whether the antiserum raised against the C-terminal fragment expressed in *E.coli* detects the 160 kDa subunit of CPSF during CPSF purification. The CPSF activity profile over a Heparin Sepharose column is shown in Figure 5A (9). When complemented with PAP, the precleaved L3 RNA substrate is specifically polyadenylated in fractions where CPSF elutes (fractions 30–66). The lane marked P contained only PAP, the one marked IP shows the activity of the Blue Sepharose pool loaded onto the column. As is clearly visible in Figure 5B, the affinity-purified antiserum recognises a protein with an apparent molecular mass of 160 kDa that coelutes with the CPSF activity. The same behaviour was observed on a DEAE column and on a Superose 6 gel filtration column (data not shown). This constitutes additional proof for the identity of the cloned cDNAs.

Sequence motifs and similarities

In contrast to the 100 kDa subunit of CPSF (21) the 160 kDa subunit contains a potential nuclear localisation signal of the bipartite type between position 894 and 909 [Fig. 3; (34)]. Two consecutive basic amino acids are followed by a spacer of nine amino acids and another three basic amino acids within the next five residues. The typical length of a spacer is 10 amino acids; however, the function of the a nuclear localisation signal is relatively insensitive to changes of the spacer distance (35). For example, a mutant construct of the influenza virus polymerase basic protein I containing a spacer sequence of only nine amino acids is still transported to the nucleus (36).

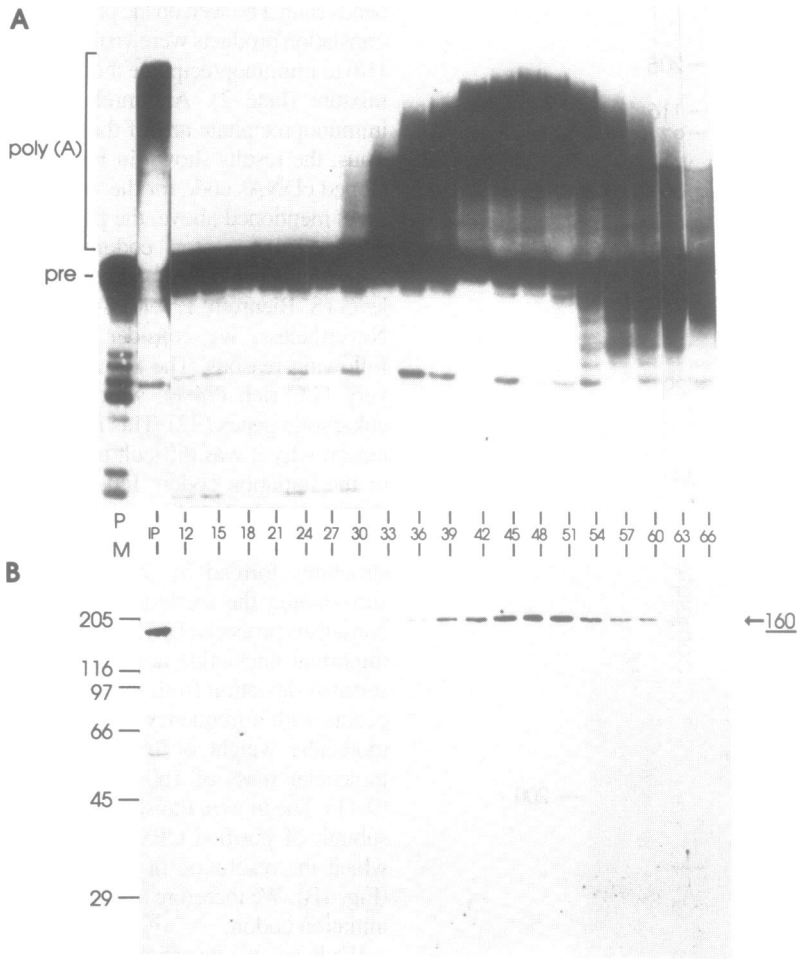

Figure 5. The 160 kDa protein recognised by the antiserum raised against the C-terminus of the cloned protein elutes together with CPSF activity from a Heparin–Sepharose column. (**A**) Specific polyadenylation of L3pre RNA in the presence of PAP with 0.5 µl of each indicated column fraction. P, PAP only; IP, 0.5 µl of the column load. On the original gel, fractions 63 and 66 were loaded in the wrong order. (**B**) The same column fractions were separated on a 9% SDS–polyacrylamide gel and blotted. The 160 kDa subunit of CPSF was detected with affinity-purified antibodies raised against the bacterially expressed C-terminus of the cDNA.

It has been shown that the 160 kDa subunit of CPSF can be crosslinked to functional polyadenylation substrates (17,21); nevertheless, no known RNA binding motif is present in the sequence (42). Neither a RNP-type RNA binding domain, nor a KH, a RGG domain, or an arginine cluster is present. Attempts to map a new type of RNA binding domain by expressing smaller protein fragments of the 160 kDa subunit in *E.coli* have failed so far because most of the fragments were insoluble (data not shown).

Searches of the EMBL and GenBank data libraries with the program TFASTA (29) revealed several human expressed sequence tags matching the 160 kDa subunit of bovine CPSF (accession numbers M78983; T60206; T59068; T19254; T19414). Except for the sequence tag M78983, that spans the region between residues 298 and 406, they all map to the region between amino acids 1066 and 1252. The identity of M78983 to the 160 kDa subunit of CPSF is 85% on the nucleotide level and 83% on the amino acid sequence level. The identities of the other tags are better than 90%. All of them contain at least one frame shift when compared to the amino acid sequence of bovine CPSF.

Searching the Swissprot database with the FASTA program (29) revealed an additional similarity with a UV-damaged DNA binding protein (UVdDb) defective in some Xeroderma pigmentosum group E patients [(37); accession number S38777]. The overall similarity is 50%. Although the overall identity is only 24%, the alignment scores of the UVdDb and the 160 kDa subunit of CPSF are significantly higher than the average score of 10 randomised sequences with the same amino acid composition. Interestingly, the C-terminal end of the UVdDb has been previously aligned to three other sequences (37): a human open reading frame 399 (38) with an incomplete N-terminus, a hypothetical protein from the slime mold *D.discoideum* (39) and an expressed sequence tag from rice (40). We have aligned the C-termini of these four proteins with the C-terminus of the CPSF 160 kDa subunit and an additional similar sequence, the open reading frame YM9827.03C on chromosome XIII of *Saccharomyces cerevisiae* (41). Figure 6 shows that, except for a 45 amino acid insertion in the yeast sequence, two clusters of homologies are detected. The second cluster comprises the last 70 amino acids of each protein. The core of this homology ([I/L]DGDL[I/L]EX[F/Y]L) is hydrophobic and contains three conserved negatively charged amino acids. Profile searches with the alignment of the last 70 amino acids did not detect any other matches, particularly no non-specific hydrophobic sequences (data not shown). The function of this conserved motif remains to be elucidated.

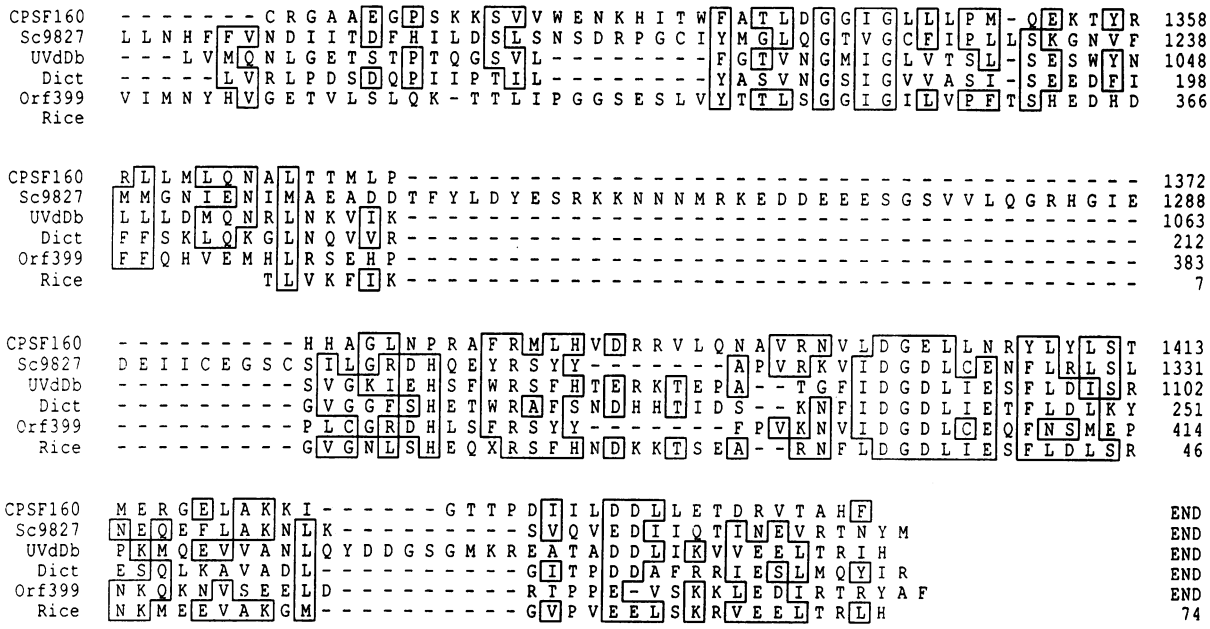

Figure 6. Alignment of the C-terminus of the CPSF 160 kDa subunit with the C-termini of *S.cerevisiae* open reading frame YM9827.03C on cosmid 9827 (Sc9827), the UV-damaged DNA binding protein from the African green monkey (UVdDb), a hypothetical protein of *D.discoideum* (Dict), the human open reading frame 399 (Orf399) and the rice sequence C0655A (Rice). The rice sequence is not known over the entire alignment.

ACKNOWLEDGEMENTS

We thank Rainer Pöhlmann for advice on DNA sequencing and Reinhard Doelz for help with the computer analysis. We are especially grateful for discussions with Elmar Wahle and Jochen Wittbrodt. Thanks for reading the manuscript go to Ursula Rüegsegger, Lionel Minvielle-Sebastia, Elmar Wahle and Mary O'Connell. This work was supported by grants fom the Kantons of Basel and the Schweizerischer Nationalfonds.

REFERENCES

1 Wahle, E. (1992) *BioEssays* **14**, 113–118.
2 Wahle, E. and Keller, W. (1992) *Annu. Rev. Biochem.* **61**, 419–440.
3 Sachs, A. and Wahle, E. (1993) *J. Biol. Chem.* **31**, 2295–2298.
4 Jackson, R. J. and Standart, N. (1990) *Cell* **62**, 15–24.
5 Decker, C. J. and Parker, R. (1993) *Genes Dev.* **7**, 1632–1643.
6 Wickens, M. (1990) *Trends Biochem. Sci.* **15**, 277–281.
7 Bardwell, V. J., Wickens, M., Bienroth, S., Keller, W., Sproat, B. S. and Lamond, A. I. (1991) *Cell* **65**, 125–133.
8 Moore, C. L. and Sharp, P. A. (1985) *Cell* **41**, 845–855.
9 Bienroth, S., Wahle, E., Suter-Crazzolara, C. and Keller, W. (1991) *J. Biol. Chem.* **266**, 19768–19776.
10 Gilmartin, G. M. and Nevins, J. R. (1989) *Genes Dev.* **3**, 2180–2189.
11 Murthy, K. G. K. and Manley, J. L. (1992) *J. Biol. Chem.* **267**, 14804–14811.
12 Takagaki, Y., Ryner, L. C. and Manley, J. L. (1989) *Genes Dev.* **3**, 1711–1724.
13 Takagaki, Y., Manley, J. L., MacDonald, C. C., Wilusz, J. and Shenk, T. (1990) *Genes Dev.* **4**, 2112–2120.
14 Christofori, G. and Keller, W. (1988) *Cell* **54**, 875–889.
15 Wahle, E., Lustig, A., Jens, P. and Maurer, P. (1993) *J. Biol. Chem.* **268**, 2937–2945.
16 Bienroth, S., Keller, W. and Wahle, E. (1993) *EMBO J.* **12**, 585–594.
17 Keller, W., Bienroth, S., Lang, K. M. and Christofori, G. (1991) *EMBO J.* **10**, 4241–4249.
18 Moore, C. L., Chen, J. and Whoriskey, J. (1988) *EMBO J.* **7**, 3159–3169.
19 Gilmartin, G. M. and Nevins, J. R. (1991) *Mol. Cell. Biol.* **11**, 2432–2438.
20 MacDonald, C. C., Wilusz, J. and Shenk, T. (1994) *J. Mol. Biol.* **14**, 6647–6654.
21 Jenny, A., Hauri, H.-P. and Keller, W. (1994) *Mol. Cell. Biol.* **14**, 8183–8190.
22 Laemmli, U. K. (1970) *Nature* **227**, 680–685.
23 Kyhse-Andersen, J. (1984) *J. Biochem. Biophys. Meth.* **10**, 203–209.
24 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
25 Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S. and Rastan, S. (1992) *Cell* **71**, 515–526.
26 Henikoff, S. (1984) *Gene* **28**, 351–359.
27 Troutt, A. B., McHeyzer-Williams, M. G., Pulendrean, B. and Nossal, G. J. V. (1992) *Proc. Natl. Acad. Sci USA* **89**, 9823–9825.
28 Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
29 Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
30 Olmstedt, J. B. (1981) *J. Biol. Chem.* **256**, 11955–11957.
31 Lathe, R. (1985) *J. Mol. Biol.* **183**, 1–12.
32 Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8148.
33 Kozak, M. (1991) *J. Biol. Chem.* **266**, 19867–19870.
34 Dingwall, C. and Laskey, R. A. (1991) *Trends Biochem. Sci.* **16**, 478–481.
35 Robbins, J., Dilworth, S. M., Laskey, R. and Dingwall, C. (1991) *Cell* **64**, 615–623.
36 Nath, S. T. and Nayak, D. P. (1990) *Mol. Cell. Biol.* **18**, 4139–4145.
37 Takao, M., Abramic, M., Moos Jr., M., Rapic' Otrin, V., Wootton, J. C., McLenigan, M., Levine, A. S. and Protic, M. (1993) *Nucleic Acids Res.* **21**, 4111–4118.
38 Nomura, N., Miyajima, N., Kawarabajashi, Y. and Tabata, S. (1993) GenBank accession number D13642.
39 Sydow, L., Alexander, H. and Alexander, S. J. (1992) EMBL accession number X65937.
40 Sasaki, T. and Minobe, Y. (1993) GenBank accession number D15449.
41 Odell, C. and Browman, S. (1995) GenBank accession number Z47816AC.
42 Burd, C. G. and Dreyfuss, G. (1994) *Science* **265**, 615–621.
43 Brosi, R., Hauri, H. and KrSmer, A. (1993) *J. Biol. Chem.* **268**, 17640–17646.