The EMBO Journal vol.5 no.12 pp.3209–3217, 1986

Cloning of the human cDNA for the U1 RNA-associated 70K protein

Hubert Theissen, Michael Etzerodt[1], Rolf Reuter, Claudio Schneider[1], Friedrich Lottspeich[2], Patrick Argos[1], Reinhard Lührmann and Lennart Philipson[1]

Max-Planck-Institut für molekulare Genetik, Otto-Warburg-Laboratorium, Ihnestrasse 73, D-1000 Berlin 33, [1]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900 Heidelberg, and [2]Max-Planck-Institut für Biochemie, Genzentrum, D-8033 Martinsried bei München, FRG

Communicated by L. Philipson

Anti-RNP sera were used to isolate a cDNA clone for the largest polypeptide of the U1 snRNP, a protein of mol. wt 70 kd designated 70K, from a human liver cDNA library constructed in the expression vector pEX1. The cro-β-galactosidase–70K fusion protein reacted with various anti-RNP patient sera, a rabbit anti-70K antiserum, as well as with a monoclonal antibody specific for this protein. The sequences of four 70K peptides were determined and they match parts of the deduced amino acid sequence of the 1.3 kb insert of p70.1 indicating that it is a genuine 70K cDNA. Screening of a new cDNA library constructed from polysomal mRNA of HeLa cells with the p70.1 clone yielded an overlapping clone, FL70K, which was 2.7 kb long and covered the complete coding and 3′-untranslated sequence of the 70K protein in addition to 680 nucleotides upstream of the putative initiation codon, The predicted mol. wt of the encoded protein is ~70 kd. Amino acid analysis of the purified HeLa 70K protein yielded values close or identical to those deduced from the nucleoide sequence of the full-length cDNA. The 70K protein is rich in arginine (20%) and acidic amino acids (18%). Extremely hydrophilic regions containing mixed-charge amino acid clusters have been identified at the carboxyl-terminal half of the protein, which may function in RNA binding. A sequence comparison with two recently cloned RNA binding proteins revealed homology with one region in the U1 RNP 70K protein. This domain may also be responsible for RNA binding.

Key words: small nuclear ribonucleoprotein/cDNA cloning/RNP antigens/systemic lupus erythematosus

Introduction

U1 RNA is the most abundant of the small nuclear RNAs of the U snRNA class which are found in all eucaryotic cells (for recent review see Brunel *et al.*, 1985). It has been proposed (Lerner and Steitz, 1979; Roger and Wall, 1980) and experimentally verified (Padgett *et al.*, 1983; Krämer *et al.*, 1984; Black *et al.*, 1985; Krainer and Maniatis, 1985) that U1 RNA is involved in pre-mRNA splicing. The participation of U1 RNA in mRNA splicing is probably mediated by base-pairing between its 5′ end and the 5′ splice site of the pre-mRNA.

U1 RNA, however, does not function as a naked RNA molecule, but exists in the nucleus as a ribonucleoprotein particle (U1 snRNP). The protein components of the U1 snRNP appear to guide the specificity of the interaction of U1 RNA with the mRNA 5′ splice sites (Mount *et al.*, 1983). At least nine polypeptides are associated with purified U1 snRNP, the polypeptides 70K, A, B, B′, C, D, E, F and G (Hinterberger *et al.*, 1983; Kinlaw *et al.*, 1983; Billings and Hoch, 1984). While the proteins B, B′, D, E, F and G are also present in the U2, U5 and U4/U6 snRNPs, the 70K, A and C polypeptides are unique to the U1 snRNP species. These polypeptides may therefore be responsible for the function of U1 snRNP in pre-mRNA splicing, supported by the fact that antibodies directed against the U1 RNP specific proteins inhibit the *in vitro* splicing reaction (Padgett *et al.*, 1983) The structure and function of these proteins have not yet been analysed, but they appear not to be derived from a common precursor polypeptide but are encoded in separate mRNAs (Fisher *et al.*, 1983; Wieben *et al.*, 1983).

The U1 snRNP-specific proteins 70K, A and C also carry the antigenic determinants for the anti-RNP autoantibodies which, like anti-Sm autoantibodies are found in sera from patients with lupus erythematosus and related disorders (Tan, 1982).

A prerequisite for an understanding of the role the RNP proteins play in splicing and in autoimmune diseases is the knowledge of their primary structure.

In the present study we report on the molecular cloning, sequencing and identification of two overlapping cDNA clones encoding the U1 RNA-associated 70K protein. The first cDNA clone was isolated from a human liver cDNA library constructed in the pEX1 expression vector (Stanley, 1983) using human anti-RNP autoantisera as a source of anti-70K antibody. After unequivocal identification, including comparison with the sequence of peptides obtained by digestion of purified 70K protein, a full-length cDNA was isolated and sequenced.

Results

Isolation and identification of a 70K-containing clone

A human liver cDNA library constructed with the expression vector pEX1 was screened using anti-RNP sera from the autoimmune patients which were positive for the U1 RNP-specific 70K polypeptide. Initial screening of 5×10^4 clones identified four putative 70K clones. After colony purification and rescreening with anti-RNP sera one positive clone, p70.1, was isolated. The p70.1 clone also reacted with a monoclonal antibody (H386) specific for the 70K protein as well as with an anti-70K rabbit antiserum. It did not react with human anti-Sm sera or with monoclonal antibodies against the Sm polypeptides.

The cro-β-galactosidase-70K fusion protein of the recombinant clone p70.1 was further characterized by immunoblot analysis. Bacteria containing this clone were grown at 42°C for 1 h to induce expression of the fusion protein, and the bacterial lysates were fractionated on a 7% SDS–polyacrylamide gel, transferred to a nitrocellulose filter and blotted to various antisera. As shown in Figure 1A (lanes 2+3), two anti-RNP sera which were used for the initial screening of the cDNA library stain a broad fusion protein band with an apparent mol. wt of ~160 000 daltons. The cro-β-galactosidase hybrid protein en-

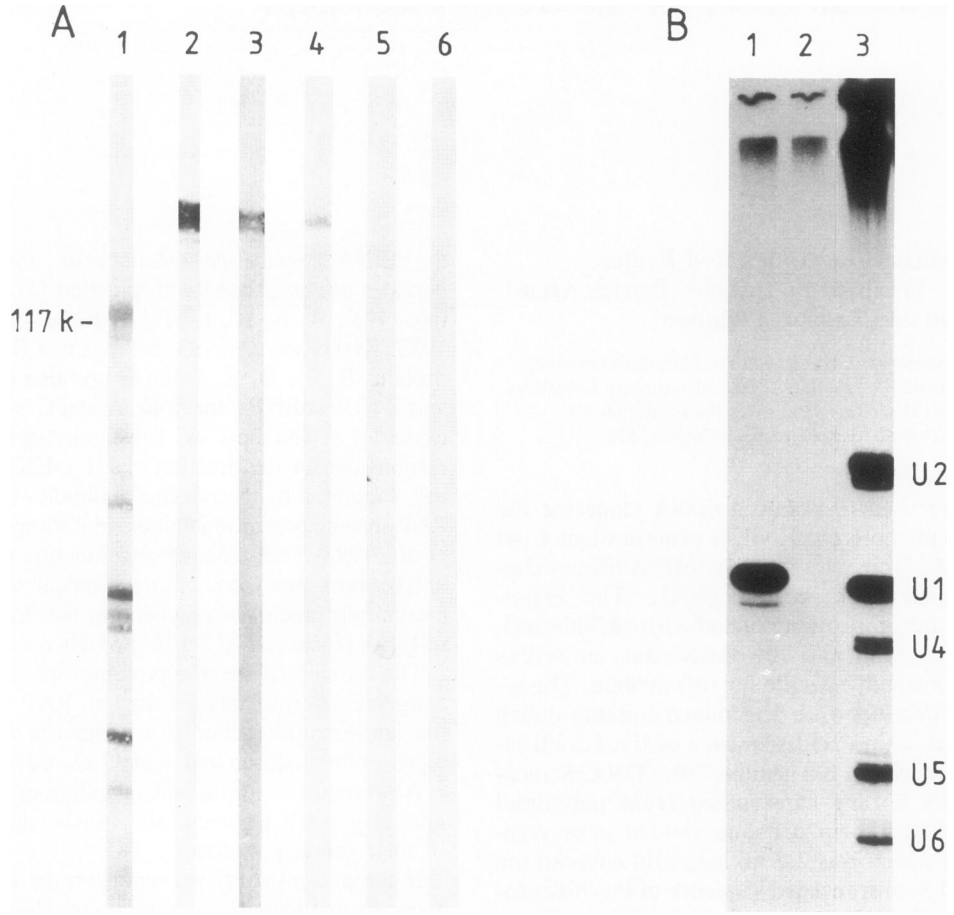


Fig. 1. Characterization of the cro-β-galactosidase-70K fusion protein by immunoblotting (**A**) and immunoprecipitation of snRNPs with affinity-purified antibodies (**B**). (**A**) Lysates were prepared from bacteria containing the p70.1 clone after growth at 42°C for 1 h and analysed by immunoblotting. Immunoblots were probed with two distinct anti-RNP sera containing antibodies against the 70K protein (**lanes 2** and **3**), with the 70K-specific monoclonal antibody H386 (**lane 4**), with an anti-Sm serum that did not contain anti-70K autoantibodies (**lane 5**) and with a normal human serum (**lane 6**). Antibody detection was with peroxidase-conjugated anti-human and anti-mouse IgG, respectively. **Lane 1** displays Coomassie-stained proteins of lysates from bacteria containing the pEX1 vector. The size of the cro-β-galactosidase hybrid protein is given in kd. (**B**) Immunoblots from bacterial lysates as described above were probed with an anti-RNP serum containing anti-70K antibodies. The region of the blot containing the cro-β-galactosidase-70K fusion protein, or a non-reactive control region from the middle of the blot were excised, bound antibodies were eluted (Smith and Fisher, 1984) and the respective fractions were reused for immunoprecipitation of snRNPs from ^{32}P-labeled HeLa cell nuclear extracts. snRNAs present in immune precipitates obtained with antibodies eluted from the fusion protein (**lane 1**) or with the control fraction (**lane 2**) were fractionated on polyacrylamide gels. **Lane 3** exhibits RNAs extracted from the nuclear extract.

coded by the pEX1 vector has a mol. wt of ~117 000 daltons (Figure 1A, lane 1).

The same protein band was also stained with the monoclonal antibody H386, which is specific for the 70K protein (Figure 1A, lane 4). Normal human serum or anti-Sm patient sera, which are negative for the 70K protein, did not stain the blots (Figure 1A, lanes 5 and 6).

To verify the specificity of the antibody reaction with the fusion protein, we prepared preparative immunoblots with an anti-RNP patient serum, re-eluted the antibodies bound to the fusion protein on the nitrocellulose filter and assayed the eluted antibody for its reactivity with intact U snRNPs. Figure 1B, lane 1, demonstrates that affinity-purified antibodies precipitated only U1 snRNPs from a nuclear extract enriched for the snRNPs U1–U6. No reaction was observed when eluted material from a control region of the nitrocellulose blots was used for immunoprecipitation of snRNPs (Figure 1B, lane 2).

These results, together with our findings that the antibodies eluted from the fusion protein band reacted strongly with the purified intact 70K protein in a microtiter ELISA, but with none of the other snRNP proteins (data not shown), indicate that the cDNA clone p70.1 contains an insert coding for a fragment of the 70K protein.

Characterization and DNA sequence of the 70K cDNA clone

The clone p70.1 contains an insert of 1.3 kb as determined by restriction mapping and sizing of the DNA fragments on agarose gels. Sequencing of the insert was performed on both strands by the Maxam and Gilbert technique. A partial restriction map of the clone and the sequencing strategy is shown in Figure 2A. The DNA sequence of the p70.1 clone is shown in Figure 3. The first 15 nucleotides constitute the *PstI* restriction site and a truncated G-tail of the cDNA insert. There is one long single open reading frame which is uninterrupted until the end of the clone and the last five nucleotides probably constitute part of the C-tail (Figure 3). Thus ~425 amino acids are encoded by this portion of the 70K protein. No putative polyadenylation signal or poly(A) stretch was found in this clone.

Limited primary sequence of the 70K protein

Direct evidence that clone p70.1 contains an insert coding for

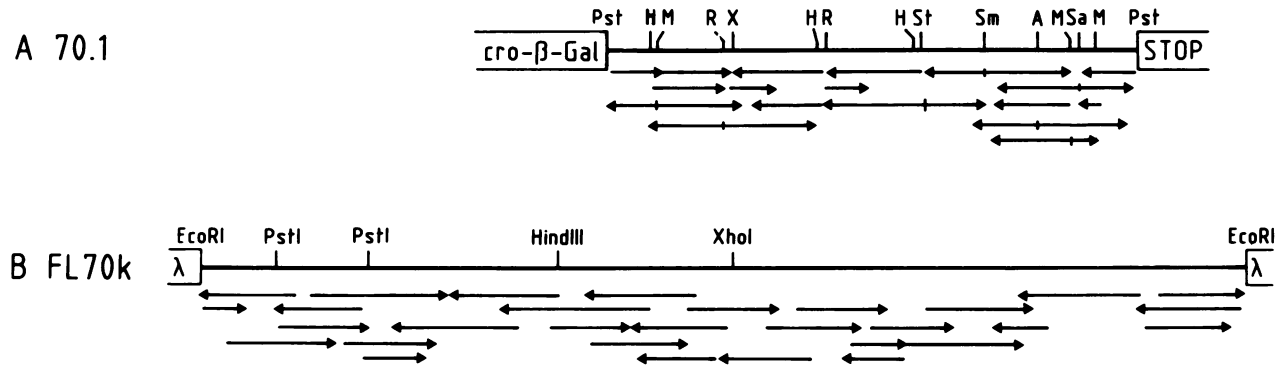


Fig. 2. Partial restriction map and sequencing strategy for cDNA clone p70.1 (**A**). Partial restriction map, identification of *Bal*31 deletion clones and sequencing strategy of the full-length cDNA clone FL70K (**B**).

the U1 snRNP protein 70K was obtained by comparing the cDNA deduced amino acid sequence with amino acid sequence analysis data from cyanogen bromide (CNBr) peptides from a 70K fragment. The protein sequencing was performed on a fragment of the 70K polypeptide structurally overlapping the cDNA encoded protein. Since treatment of the 70K protein with trypsin generates a 40 kd fragment which is reactive with anti-RNP antibodies (White *et al.*, 1982), we incubated purified snRNPs U1–U6 with limiting amounts of trypsin to produce large amounts of the 40 kd fragment from U1 RNA-associated 70K polypeptide. Following preparative SDS–PAGE of the total protein mixture, the 40 K fragment was eluted electrophoretically from the gels. Comparative peptide mapping of the isolated 40 kd fragment and the 70K protein using V8 protease and chymotrypsin confirmed that the former was a fragment of the 70K protein (data not shown). A structural overlap of the 40 kd fragment with the p70.1 cDNA protein was verified by showing that monoclonal antibody H386 and the anti-RNP sera used to identify the p70.1 cDNA clone also reacted with the 40 kd fragment on immunoblots (not shown).

It was not possible to sequence the amino-terminus of the 40 kd fragment, probably due to a blocked N-terminal amino acid generated during the isolation. The 40 kd fragment was therefore cleaved with CNBr and the mixture of the resulting peptides was sequenced on a gas-phase sequencer. Four distinct phenylthiohydantoin amino acids could be separated by h.p.l.c. after each cycle, indicating that four peptides containing unblocked amino termini had been generated from the 40 kd polypeptide by CNBr cleavage. In general, the yields were sufficient for unambiguous indentification of all amino acids during nine cycles (Table I). Only lysine residues were recovered in low yield, probably due to the destruction of the lysine side chain during the CNBr cleavage. Most important, the four amino acid residues of each cycle can be aligned to match the sequence of four 9-mer peptides in the cDNA clone p70.1 (Figure 3). Furthermore, the four CNBr peptides are preceded by methionine in the cDNA sequence (Figure 3).

Table I. Sequence analysis of a mixture of CNBr peptides derived from a 40 kd fragment[a] of the 70K protein

Cycle	Amino acids[b]			
1	E	W	V	H
2	R	D	Y	S
3	(K)[c]	P	S	A
4	R	H	(K)	Y
5	R	N	R	(K)
6	E	D	S	H
7	(K)	P	G	A
8	I	N	(K)	D
9	E	A	P	G

[a]Isolation of the 40-kd fragment and sequence analysis of the CNBR fragments were performed as described in Results and Materials and methods.
[b]Amino acids are given in the one-letter code.
[c]Lysine residues were recovered in low yields and could therefore only be tentatively assigned (K).

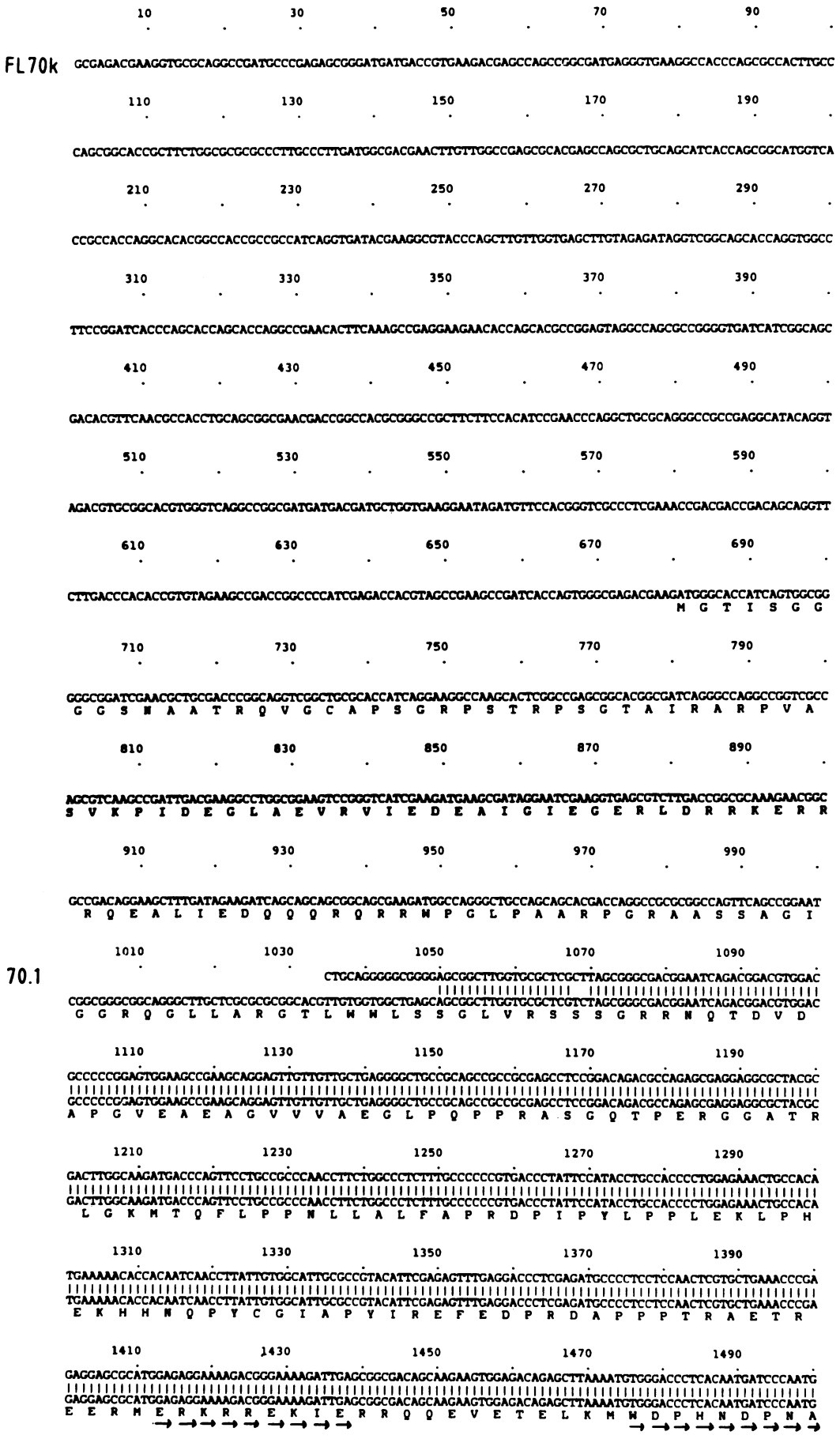
Isolation and sequence of a full-length 70 K cDNA

These results led us to isolate a full-length cDNA clone taking advantage of a novel method based on affinity purification of the first strand cDNA complex on anti-m7G antibodies followed by RNase A digestion and elution with m7G before second strand synthesis (Schneider *et al.*, 1986). The clones were inserted with *Eco*RI linkers in the vector λ1149 (Murray, 1983) and screened for cross-hybridization with the nick-translated *Pst* fragment of the p70.1 clone. A single clone (FL70) among 10^5 hybridized with the probe and this phage contained an insert of 2.7 kb. Restriction enzyme maps and Southern blots established that it contained the p70.1 sequence and it was subcloned in pEMBL19. Several *Bal*31 clones were made in the M13, mp18 and mp19 vectors, as shown in Figure 2B and all clones were sequenced by the Sanger technique establishing the sequence shown in Figure 3. Analysis of this clone shows a long open reading frame spanning nucleotide 681–2523 (Figure 3). Most important, this reading frame contains a precise match with the four CNBr peptide sequences which we deduced from direct sequence analysis of the 40 kd fragment of the 70K protein as described above (Figure 3). A polyadenylation signal and the polyadenylation site are also present in this clone in addition to 680 nucleotides upstream from the putative translation initiation site (Figure 3).

Amino acid analysis of the purified 70K protein yields values close or identical to those derived from the nucleotide sequence of the long open reading frame of the FL70K clone (Table II).

RNA blot hybridization of 70K mRNA

To investigate the corresponding RNA species, poly(A)$^+$ RNA from HeLa cells and total RNA from a human hepatoma cell line (HEPG2) were fractionated in a denaturing formaldehyde/agarose gel, blotted onto nitrocellulose paper and subse-


FL70k
70.1

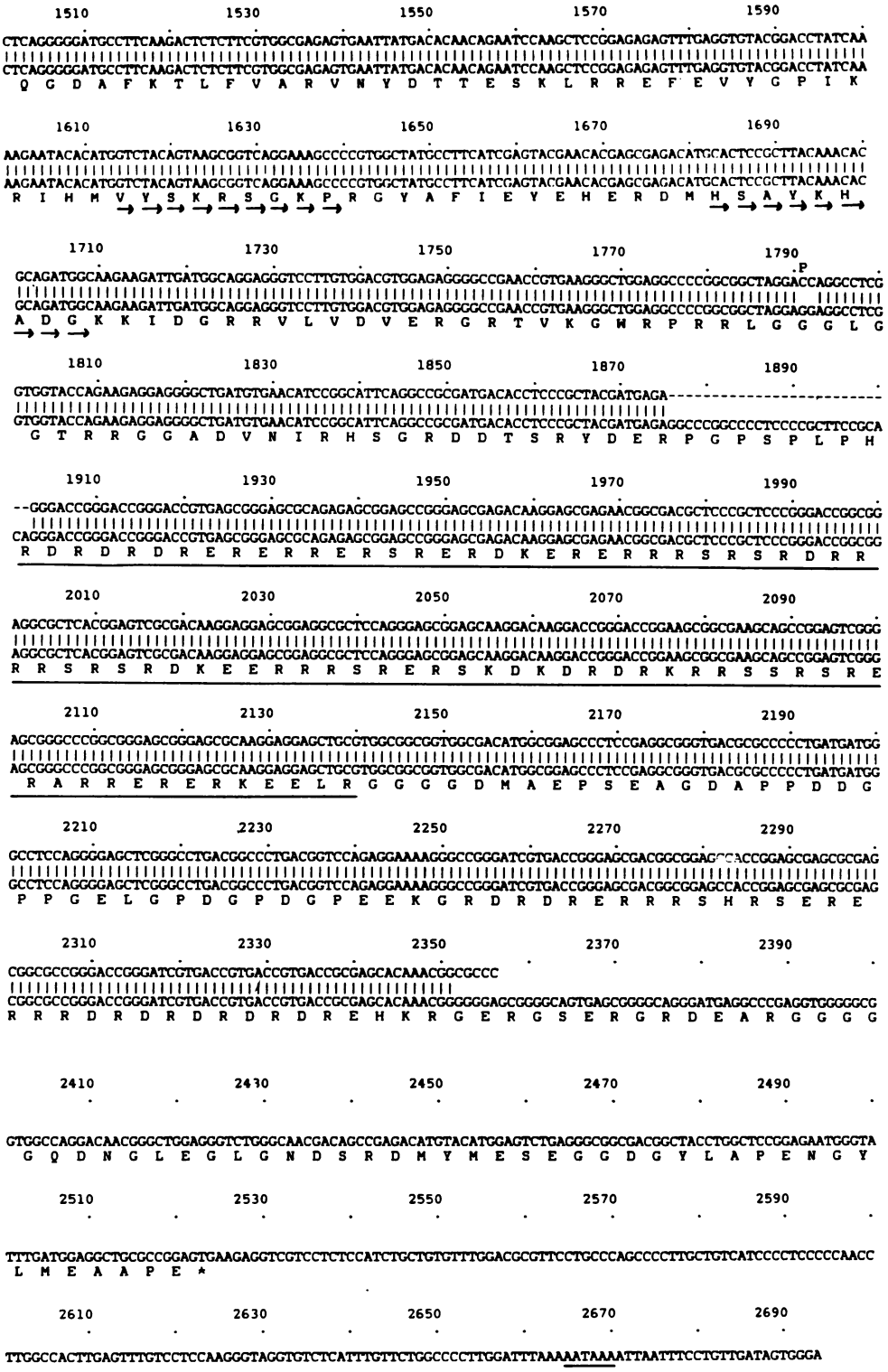


Fig. 3. DNA sequence of the two cDNA clones and the deduced amino acid sequence of the encoded 70K protein. The deduced amino acid sequence is displayed below the DNA sequence (in the one-letter code). The sequences of the four cDNA-derived 9-mer peptide matching the respective 9-mer peptide sequences which can be deduced from the direct amino acid sequence anlaysis of a CNBr peptide mixture of the 40 K fragment are displayed below the DNA-deduced sequence. Lysine residues (→) could only be tentatively assigned due to the partial destruction of lysine side chains during the CNBr cleavage of the 40 K fragment. The long mixed-charge amino acid cluster is underlined. The poly(A) addition signal is also underlined and the poly(A) addition site is at the end of the sequence. Vertical lines designate identity between the nucleotide sequence of p70.1 and pFL70K. Horizontal lines indicate gaps introduced into the sequence of p70.1 in the alignment for maximum overlap with FL70K.

quently hybridized with the two nick-translated cDNA clones. The cDNA plasmid p70.1 hybridized strongly to an RNA molecule of ~1700 nucleotides from HeLa cells and ~3000 nucleotides from hepatoma cells (Figure 4). The same results were obtained with the FL70 clone. The larger mRNA was also observed in HeLa cells although it gave a weaker signal. The relationship between these two RNA species will be reported separately.

Table II. Amino acid composition of cloned and isolated 70 K proteins

Amino acid	FL70K[a]		Isolated 70K[b]
Asp	48	59	54
Asn	11		
Glu	66	82	82
Gln	16		
Ser	36		36
Gly	71		66
Thr	18		18
His	12		14
Ala	41		40
Arg	117		116
Tyr	13		18
Val	21		22
Met	10		7
Ile	17		18
Phe	7		12
Leu	32		34
Lys	25		23
Cys	2		ND[c]
Pro	46		ND
Trp	5		ND

[a]Results are expressed as amino acid residues/70K protein. FL70K data are deduced from the nucleotide sequence of pFL70K.
[b]The data for the isolated 70K protein represent the mean values from three independent experiments.
[c]Not determined.

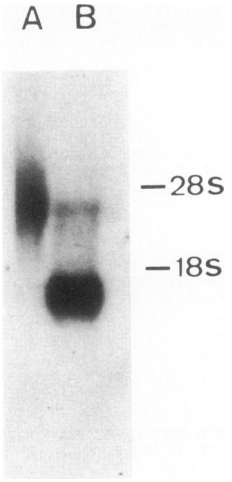


Fig. 4. Northern blot analysis of total RNA from a hepatoma cell line (HEPG2; **lane A**) and of poly(A)$^+$ RNA from HeLa cells (**lane B**) probed with a ^{32}P-labeled nick-translated cDNA insert from the p70.1 clone, as described in Materials and methods. The size of concurrently electrophoresed ribosomal RNA from HeLa cells is also indicated.

Discussion

We isolated and sequenced two human cDNA clones for the U1 snRNA-associated 70K protein. Several pieces of evidence indicate that the clones are authentic 70K cDNAs: (i) the polypeptide fragment expressed as a cro-β-galactosidase fusion protein in the pEX vector reacts with various human anti-RNP sera, while it did not react with anti-Sm antibodies. Furthermore, the fusion protein was recognized by a monoclonal antibody specific for the 70K protein; (ii) antibodies eluted from the fusion protein band on immunoblots selectively precipitated native U1 snRNPs but none of the other U snRNPs; (iii) most convincingly, the DNA-deduced amino acid sequence matches four 9-mer peptide sequences which can be deduced from direct amino acid sequence analysis of a CNBr peptide mixture from a fragment of the 70K protein. We are therefore confident that we have correctly identified the reading frame of the 70.1 cDNA clone.

Screening of a HeLa cDNA library with the p70.1 clone yielded an overlapping clone, FL70K, which covered the complete coding and 3'-untranslated sequence of the 70K protein in addition to 680 nucleotides upstream of the putative initiation codon (Figure 3). The long open reading frame of this clone codes for 614 amino acids. The predicted mol. wt of this protein is ~72 000 which is in good agreement with the apparent mol. wt of ~68–75 kd, as estimated previously by SDS–PAGE (Hinterberger *et al.*, 1983; Bringmann *et al.*, 1983b; Billings and Hoch, 1983). The amino acid composition of the 70K protein deduced from the nucleotide sequence of FL70K is compared to the experimentally determined one of the isolated HeLA 70K protein in Table II. The values are either close to each other or identical which further supports our notion that we have isolated a full-length cDNA from the 70K protein.

When the nucleotide sequences of the two isolated 70K encoding cDNAs are aligned, they are identical except for three positions (Figure 3). A TCT→CTT change is found around position 1072 and a GGA→CCA change around position 1792. While in the first case both triplets code for the same amino acid, in the second case glycine switches to proline in p70.1 (Figure 3). A third difference between the two cDNA sequences is found for the region spanning nucleotides 1876–1902. These 27 nucleotides are only present in pFL70K but not in p70.1 (Figure 3). In view of the fact that the two cDNAs were isolated from distinct cDNA libraries which had been constructed using mRNA from HeLa cells and human liver, respectively, the observed differences could indicate that the 70K protein is polymorphic at the DNA and amino acid level. The analysis of human genomic 70K clones will provide further insight into this question.

One finding which requires a comment is the discrepancy between the size of the mRNA revealed by Northern blots and the size of the full-length cDNA after cloning. The 3.0 kb mRNA was the dominant species in hepatoma cells but HeLa cells contained a major species around 1.7 kb and a minor one around 3.0 kb. The possibility that the larger cDNA may have been derived from a nuclear precursor RNA is not likely since the cDNA library from which FL70K was recovered was constructed from polysomal mRNA (Schneider *et al.*, 1986). Furthermore, the 1.7 kb mRNA codes probably only for a protein of ~55 kd which does not correspond to the observed mol. wt of this polypeptide. Since RNA binding proteins obviously share some sequence homology (see below) a separate study will be carried out to determine the relationship between the two mRNA species in HeLa cells.

The amino acid composition of the 70K protein deduced from the FL70K DNA clone is noteworthy for its high content in basic (23%) and acidic (18%) amino acid residues (Table II). Arginine is particularly enriched in the 70K protein (~20% of all amino acids). Analysis of the primary sequence for distribution of hydrophobicity (not shown) reveals only few distinctly hydrophobic stretches such as the amino acids encoded by nucleotides 1017–1049 and 1233–1253, respectively (Figure 3). Most of the remainder of the 70K protein is highly hydrophilic, which is particularly true of the carboxy terminal half. The region spanning the amino acids encoded by nucleotides 1902–2141 shows extreme hydrophilicity values and is further exceptional in that it is composed of only six distinct amino acids, among which

```
70K        RERRRSRSRDRRRRSRSRDKEERRRSR
PROTAMINE  RRRRRSSRPVRRRRRYRRSTVARRRRR
```

Fig. 5. Comparison of the spacing of oligoarginine clusters found in protamine Thyurin Z1 (Bretzel, 1973) and in the hydrophilic region of the 70K protein spanning the amino acids encoded by nucleotides 1965–2045 (see Figure 3). The figure displays only the Thyurin Z1 sequence spanning amino acids 2–28.

arginine (50%), glutamic and aspartic acids (25%) and serine (15%) dominate (Figure 3).

Within the stretch of amino acids encoded by nucleotides 1965–2045 most of the arginines are clustered in sequences of three to four which are periodically spaced. Interspersed among these clusters are either serines or acidic amino acids. A periodic spacing of arginine clusters in sequences of four to six is characteristic of protamines. In Figure 5 the sequence of the oligoarginine-rich 70K region is compared with the sequence of protamine Thyurin Z1. It is noteworthy that the spacing of the arginine clusters is similar in both proteins. This similarity is of interest since protamines are known to form tight complexes with nucleic acids (Warrant and Kim, 1978). In particular, it has been inferred from crystallographic studies of tRNA–protamine complexes that the guanidinium groups of the short α-helical oligoarginine stretches appear to interact with tRNA phosphates (Warrant and Kim, 1978).

More interesting is that Sachs *et al.* (1986) have recently determined the nucleotide sequence of a yeast poly(A)-binding protein (PABP). The PABP amino acid sequence shows four repeating units, each ~90 amino acid residues long, which are suggested to be poly(A)-binding domains. In the U1 RNP 70K protein, whose sequence is presented here, one such unit is easily observed to be homologous to the PABP domains. Cobianchi *et al.* (1986) have recently determined the nucleotide sequences of a rat helix-destabilizing protein (HDP), showing considerable sequence homology to the A1 core protein of the HnRNP (Kumar *et al.*, 1986). Although they observed no repeating units in the HDP primary structure, it is obvious that two sequence segments are related to the aforementioned spans. The alignment of the seven regions is shown in Figure 6 along with the average secondary structure prediction (Argos, 1985) and alignment positions where residues are strongly conserved.

Since the PABP and HDP proteins are known to bind single-stranded RNA or DNA, each domain may be responsible for binding nucleic acid. The passive nature of the completely conserved residues (generally hydrophobic amino acids and glycine) would not support any enzymatic function for the proteins. The secondary structure predictions indicate three protein-surface exposed regions (α-helices and turns in Figure 6) containing conserved residues that could associate with nucleic acid. Region 1 displays the strongest conservation with one position maintained acidic (E, D, N); region 2 contains many basic residues for possible phosphate interaction on the nucleotide backbone. It is possible that the spans act in concert to bind the nucleic acid or bind at different sites. Discovery of such repeats in other proteins should delineate more clearly the essential regions; however, a search of the present sequence data banks yielded no residue patterns similar to those found here. Nevertheless, such a pattern may frequently occur in RNP proteins, the majority of which have not yet been cloned.

Mean correlation coefficients were calculated; over six physical characteristics important in protein folding, for the aligned residues in each unit and for all possible pairwise comparisons

```
PABP1   A T G S E S Q S V E N S S A S L Y V G D
PABP2   W S Q R D P S L R K K G S G N I F I K N
PABP3   E R D S Q L E E T K A H Y T N L Y V K N
PABP4   E A Y R L E K M A K Y Q G V N L F V K N
HDP1    M S K S E S P K E P E Q L R K L F I G G
HDP2    - - D S Q R P G A H L T V K K I F V G G
FL70    D P H N D P N A Q G D A F K T L F V A R
        c t t t t t t c c t t t t b b b b b t
                  :                 + + +

PABP1   L E P S V S E A H L Y D I F S P I G S V
PABP2   L H P D I D N K A L Y D T F S V F G D I
PABP3   I N S E T T D E Q F Q E L F A K F G P I
PABP4   L D D S V D D E K L E E E F A P Y G T I
HDP1    L S F E T T D E S L R S H F E Q W G T L
HDP2    I K E D T E E H H L R D Y F E Q Y G K I
FL70    V N Y D T T E S K L R R E F E V Y G P I
        t t t t a a a a a a a a a a t t t t b
        +           + +       * : * +
                    1 1 1 1 1 1 1 1 1 1

PABP1   S S I R V C R D A I T K T S L G Y A Y V
PABP2   L S S K I A T D E N - G K S K G F G F V
PABP3   V S A S L E K D A D - G K L K G F G F V
PABP4   T S A K V M R T E N - G K S K G F G F V
HDP1    T D C V V M R D P N T K R S R G F G F V
HDP2    E V I E I M T D R G S G K K R G F A F V
FL70    K R I H M V Y S K R S G K P R G Y A F I
        b b b b b b t t t t t t t t t t t b b
          :   + +                 :   : * + + +
                                  2 2 2 2 2

PABP1   N F N D H E A G R K A I E Q L N Y T P I
PABP2   H F E E E G A A K E A I D A L N G M L L
PABP3   N Y E K H E D A V K A V E A L N D S E L
PABP4   C F S T P E E A T K A I T E K N Q Q I V
HDP1    T Y A T V E E V D A A M N A R - P H K V
HDP2    T F D D H D S V D K I V I Q K - Y H T V
FL70    E Y E H E R D M H S A Y K H A D G K K I
        b b c c c a a a a a a a a a c c b b b
          +                 : + :
                            3 3 3 3 3

PABP1   K G R L C R I M - - - - - - - - -
PABP2   N G Q E I Y V A P H L S R K - - -
PABP3   N G E K L Y V G R A Q K N E R M
PABP4   A G K P L Y V A I A Q R K D V R R
HDP1    D G R V V E P K R A V S R E - - -
HDP2    N G H N C E V R K A L S K Q E M A
FL70    D G R R V L V D V E R G R T V K G
        t t t b b b b c c c c c c c c c
          *     + :                 :
```

Fig. 6. Repeating domain sequences found in poly(A)-binding protein (PABP1 to PABP4) (Sachs *et al.*, 1986), helix-destabilizing protein (HDP1 and HDP2) (Cobianchi *et al.*, 1986), and the U1 RNP 70K protein (FL70). The average secondary structure prediction (Argos, 1985) is also shown where a is used to indicate helices, b for strand structure, t for turn, and c for coil. An alignment position is marked (*) if all the residues are identical; (+) if they are all conserved according to the groupings (PG), (ST), (EDQN), (KR), (IAVLMC), and (FYWH); and (:) if six or seven are conserved.

of the seven units. Argos (1985) has described the procedure. PABP1 is most closely related to PABP2 at a mean correlation of 0.50; similarly PAPB2 is most related to PABP3 (0.51); and PABP3 to PAPB4 (0.48). These results would suggest that each genomic unit may be duplicated from the existing 5′-most domain. Similarly, HDP2 is closest to HDP1 (0.47). Furthermore, since the U1 RNP 70K polypeptide contains only one such domain, it may insert itself as a single unit in genetic elements.

It is, therefore, tempting to suggest that the oligoarginine-rich region or this putative domain in the 70K protein may represent RNA binding domains. It could either be necessary for the binding of the 70K protein to U1 RNA itself, or for the proper alignment of the 5′ end of U1 RNA to the 5′ splice site of an mRNA precursor molecule during splicing. While intermolecular base-pairing between the two RNAs is probably the dominant feature of this interaction, it is none the less clear that U1 RNP proteins

are also needed for this interaction (Mount *et al.*, 1983). Using the cDNA clone for expression of appropriate protein fragments of the 70K protein, direct studies can now be performed to establish whether fragments interact with U1 RNA or defined mRNA sequences.

An interesting feature of the FL70K cDNA is the occurrence of various AUG codons within the first 680 nucleotides, some of which are in-frame with the long open reading frame downstream (Figure 3). As they are followed by in-frame termination codons they cannot function as initiation codons for the 70K protein. While in most eucaryotic mRNAs the 5'-proximal AUG is the initiation codon for protein synthesis there is a significant number of animal cell mRNAs where this is not found to be true (Kozak, 1984), showing that internal AUGs can be recognized by eucaryotic ribosomes. The insertion of upstream AUGs can reduce the efficiency of initiation downstream, however (Kozak, 1983). In this respect it will be of interest to investigate whether the 5'-terminal region of the 70K mRNA plays a role with regard to the regulation of expression of the 70K protein.

In addition to the application of the cDNAs for the study of the function and expression of the 70K protein the availability of a full-length cDNA provides a means for a detailed study of the antigenic structure of the 70K polypeptide and especially of the epitopes recognized by anti-RNP autoantibodies.

Materials and methods

Antisera and monoclonal antibodies against the 70K protein

Three separate anti-RNP autoantisera from patients diagnosed as having mixed connective tissue disease (MCTD) were used for the initial screening of the pEX cDNA library. The antisera were demonstrated to contain antibodies against the U1 RNP-specific protein 70K by immunoblots on total proteins of purified snRNPs U1–U6. A monoclonal antibody (H386) specific for the 70K protein was derived from a mouse immunized with purified U1 RNP as described elsewhere (R.Reuter and R.Lührmann, in press). A rabbit antiserum specific for the 70K protein was raised by immunization with the purified 70K protein which was electroeluted from preparative SDS–polyacrylamide gels.

Radioimmunoprecipitation and immunoblotting

Growth of HeLa cells in suspension culture, labeling of the cells with [^{32}P]orthophosphate, and preparation of nuclear extracts was performed essentially as described previously (Bringmann *et al.*, 1983a). Precipitation of antigen–antibody complexes by Protein A–Sepharose was carried out as described by Matter *et al.* (1982). Immunoblotting was performed according to the procedure of Towbin *et al.* (1979), and antigenic proteins were visualized on nitrocellose strips as described by Habets *et al.* (1983). Purification of antibodies from nitrocellulose blots was performed as described by Smith and Fisher (1984).

Isolation of snRNPs and snRNP proteins

snRNPs U1–U6 were purified from nuclear extracts of HeLa S3 cells by immunoaffinity chromatography with rabbit anti-m_3G IgG essentially as described previously (Bringmann *et al.*, 1983b). For the preparation of single snRNP polypeptides the protein mixture of the purified snRNPs was separated by electrophoresis on SDS–polyacrylamide gels. After displaying the protein bands by staining with Coomassie Blue, gel slices containing single proteins were cut out of the gel and the proteins recovered by electroelution according to the procedures described by Hunkapiller *et al.* (1983).

Screening of expression libraries with antibody probes

Human liver cDNA libraries constructed with pEX1 were kindly provided by K.Stanley (EMBL, Heidelberg). A total of 5×10^4 clones have been screened by the colony-blot procedure using human anti-RNP autoantisera, essentially as described by Stanley and Luzio (1984). For detection of antigen–antibody complexes rabbit anti-human IgG (Dianova, Hamburg) coupled to horseradish peroxidase was used.

DNA sequencing

*Pst*I inserts isolated from pEX1-70K clones were fragmented using restriction enzymes, and fragments were purified by agarose gel electrophoresis. Appropriate DNA fragments were ^{32}P-labeled using polynucleotide kinase, and subjected to chemical sequence analysis following the protocol of Maxam and Gilbert (1982).

Screening of the HeLa cDNA library, subcloning and dideoxy-sequencing were carried out according to standard methods (Maniatis *et al.*, 1982; Sanger *et al.*, 1980; Messing, 1983). Gene screen membrane was purchased from New England Nuclear. The restriction enzymes *PSt*I, *Sac*I, *Kpn*I and the Klenow fragment of DNA polymerase I were from Boehringer Mannheim. *Eco*RI, *Alu*I, *Rsa*I and exonuclease *Bal*31 were from New England Biolabs; [α-^{32}P]- and [α-^{35}S]dATP were purchased from Amersham.

The obtained nucleotide sequences were fused and analysed with software developed by Staden (1980) and Devereux and Haeberli (1984).

RNA blot hybridization

Total RNA was isolated from HeLa cells by the method of Chirgwin *et al.* (1979). Poly(A)$^+$ RNA prepared from total HeLa cellular RNA by repeated chromatography on oligo (dT)–cellulose (Maniatis *et al.*, 1982). Poly(A)$^+$ RNA was resolved by electrophoresis on a 1% agarose gel in the presence of formaldehyde (Lehrach *et al.*, 1977), and the nitrocellulose blot was prepared as described (Maniatis *et al.*, 1982). The blot was probed with ^{32}P-labeled nick-translated cDNA clone p70.1. Hybridizations were carried out at 42°C in buffer containing 50% formamide, 5 × SSC, 20 mM Na-phosphate, pH 6.5, 10 × Denhardt's solution and 10% dextran sulfate. The blot was washed twice for 20 min at room temperature in 2 × SSC and twice for 20 min at 55°C in 0.2 × SSC (Maniatis *et al.*, 1982).

Amino acid sequence analysis

N-terminal amino acid sequence analysis were performed on a gas-phase sequencer 470 A from Applied Biosystems. The phenylthiohydantoin amino acid derivatives were analysed by an h.p.l.c. system which separates all components isocratically (Lottspeich, 1985).

Amino acid analysis

0.1–0.5 μg of peptide material was hydrolysed by gas-phase hydrolysis using a trifluoroacetic acid/hydrochloric acid mixture at 159°C for 20 min (A.Tsugita and W.Mewes, submitted). The amino acid mixture was derivatized with *o*-phtaldialdehyde/mercaptopropionic acid and subjected to reversed-phase h.p.l.c.

Acknowledgements

The expert technical assistance of Susanne Rothe, Christine Anderson and Robin Buckland is gratefully acknowledged. We would also like to thank Drs Riccardo Cortese, Bernhard Dobberstein, Iain Mattaj and Keith Stanley for advice and help at different phases of this study. This work was supported by a grant from the Deutsche Forschungsgemeinschaft to R.L. (Lu 294/2-1). H.T. was supported by a short-term fellowship from EMBO and M.E. by a fellowship from the Danish Natural Science Research Council (11-5668).

References

Argos,P. (1985) *EMBO J.*, **4**, 1351–1355.
Billings,P.B. and Hoch,S.O. (1983) *J. Immunol.*, **131**, 347–351.
Black,D.L., Chabot,B. and Steitz,J.A. (1985) *Cell*, **42**, 737–750.
Bretzel,G. (1973) *Hoppe Seyler's Z. Physiol. Chem.*, **354**, 312–320.
Bringmann,P., Reuter,R., Rinke,J., Appel,B., Bald,R. and Lührmann,R. (1983a) *J. Biol. Chem.*, **258**, 2745–2747.
Bringmann,P., Rinke,J., Appel,B., Reuter,R. and Lührmann,R. (1983b) *EMBO J.*, **2**, 1129–1135.
Brunel,C., Sri-Widada,J. and Jeanteur,P. (1985) In Hahn,F.E. (ed), *Progress in Molecular and Subcellular Biology*. Springer-Verlag, Vol. 9, pp. 1–52.
Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Rutter,W.J (1979) *Biochemistry*, **18**, 5294–5299.
Cobianchi,F., Sen Gupta,D.N., Zmudzka,B.Z. and Wilson,S.H. (1986) *J. Biol. Chem.*, **261**, 3521–3543.
Devereux,J. and Haeberli,P. (1984) Program Library of the University of Wisconsin Genetics Computer Group.
Fisher,D.E., Conner,G.E., Reeves,W.H., Blobel,G. and Kunkel,H.G. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 6356–6360.
Habets,W.J., De Rooij,D.J., Saldin,M.H., Verhagen,A.P., van Eekelen,C.A., van de Putte,L.B. and van Venrooij,W.J. (1983) *Clin. Exp. Immunol.*, **54**, 265–276.
Hinterberger,M., Pettersson,I. and Steitz,J.A. (1983) *J. Biol. Chem.*, **258**, 2604–2613.
Hunkapiller,M.W., Lujan,E., Ostrander,F. and Hood,L.E. (1983) *Methods Enzymol.*, **91**, 227–236.
Kinlaw,C.S., Robberson,B.L. and Berget,S.M. (1983) *J. Biol. Chem.*, **258**, 7181–7189.
Kozak,M. (1983) *Microbiol. Rev.*, **47**, 1–45.
Kozak,M. (1984) *Nucleic Acids Res.*, **12**, 857–872.
Krainer,A.R. and Maniatis,T. (1985) *Cell*, **42**, 725–736.
Krämer,A., Keller,W., Appel,B. and Lührmann,R. (1984) *Cell*, **36**, 299–307.
Kumar,A., Williams,K.R. and Szer,W. (1986) *J. Biol. Chem.*, **261**, 11266–11273.

Lehrach,H., Diamond,D., Wozney,J.M. and Boedtker,H. (1977) *Biochemistry*, **16**, 4743–4751.
Lerner,M.R. and Steitz,J.A. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 5495–5499.
Lottspeich,F. (1985) *J. Chromatogr.*, **326**, 321–327.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Matter,C., Schapfer,U., Wilhelm,J.A., Nyffenegger,T., Parisot,R.F. and de Robertis,E.M. (1982) *Arthritis Rheum.*, **25**, 1278–1283.
Maxam,A.M. and Gilbert,W.C. (1980) *Methods Enzymol.*, **65**, 499–560.
Messing,J. (1983) *Methods Enzymol.*, **101**, 20–78.
Mount,S.M., Pettersson,I., Hinterberger,M., Karmas,A. and Steitz,J.A. (1983) *Cell*, **33**, 509–518.
Murray,N. (1983) In Hendrix,R.W., Roberts,J.W., Stahl,F.W. and Weisberg, R.A. (eds), *Lambda II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 395–432.
Padgett,R.A., Mount,S.M., Steitz,J.A. and Sharp,P.A. (1983) *Cell*, **35**, 101–107.
Rogers,J. and Wall,R. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 1877–1879.
Sachs,A.B., Bond,M.W. and Kornberg,R.D. (1986) *Cell*, **45**, 827–835.
Sanger,F., Coulson,A.R., Barral,B.G., Smith,A.J.H. and Roe,B.A. (1980) *J. Mol. Biol.*, **143**, 161–178.
Smith,D.E. and Fisher,P.A. (1984) *J. Cell Biol.*, **99**, 20–28.
Staden,R. (1980) *Nucleic Acids Res.*, **8**, 3673–3694.
Stanley,K.K. (1983) *Nucleic Acids Res.*, **11**, 4077–4092.
Stanley,K.K. and Luzio,J.P. (1984) *EMBO J.*, **3**, 1429–1434.
Tan,E.M. (1982) *Adv. Immunol.*, **33**, 167–240.
Towbin,H., Staehelin,T. and Gordon,J. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 4350–4354.
Warrant,R.W. and Kim,S.H. (1978) *Nature*, **271**, 130–135.
White,P.J., Billings,P.B. and Hoch,S.O. (1982) *J. Immunol.*, **128**, 2751–2756.
Wieben,E.D., Madore,S.J. and Pederson,T. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1217–1220.

Received on 18 August 1986; revised on 22 September 1986