Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 9669–9673, December 1988
Genetics

Isolation and characterization of a *Xenopus laevis* C protein cDNA: Structure and expression of a heterogeneous nuclear ribonucleoprotein core protein

FRANK PREUGSCHAT AND BARBARA WOLD*

Division of Biology, California Institute of Technology, Pasadena, CA 91125

Communicated by John Abelson, May 31, 1988

ABSTRACT The C proteins are major components of heterogeneous nuclear ribonucleoprotein complexes in nuclei of vertebrate cells. To begin to describe their structure, expression, and function we isolated and determined the DNA sequence of *Xenopus laevis* C protein cDNA clones. The protein predicted from the DNA sequence has a molecular mass of 30,916 kDa and is very similar to its human counterpart. Although mammalian genomes contain many copies of C protein sequence, the *Xenopus* genome contains few copies. When C protein RNA was synthesized *in vitro* and microinjected into stage-VI *Xenopus* oocytes, newly synthesized C proteins were efficiently localized in the nucleus. *In vitro* rabbit reticulocyte lysate and *in vivo Xenopus* oocyte translation systems both produce from a single mRNA two discrete polypeptide species that accumulate in a ratio similar to that of mammalian C1 and C2 proteins *in vivo.*

Biogenesis of mRNA in eukaryotes features a complex series of posttranscriptional events. These include processing and modification of primary transcripts to remove intervening sequences and create 3′ termini, selective degradation of a large fraction of newly synthesized heterogeneous nuclear RNA (hnRNA), and efficient translocation of mature mRNA to the cytoplasm. These critical events in RNA metabolism display high sequence specificity and can be regulated both qualitatively and quantitatively to yield biologically significant alterations in the pattern of gene expression. Several diverse lines of evidence indicate that the substrates for intranuclear RNA processing and transport are not naked nucleic acids, but rather are ribonucleoprotein (RNP) complexes. These nuclear RNP complexes are subsequently dismantled in conjunction with export of RNA to the cytoplasm where messages are associated with a new set of RNPs (for reviews, see refs. 1 and 2).

The first RNP complexes to form on polymerase II transcripts are referred to as heterogeneous nuclear RNPs (hnRNPs) or ribonucleosomes, and they are assembled cotranscriptionally. These have been visualized in striking electron micrographs in which nascent transcripts are seen in a pattern of 200-Å RNP "beads on a string" (3–7). Limited ribonuclease digestion of nuclear RNP yields a population of relatively uniform stable monomers that sediment at 30–40S under temperate salt and pH conditions (8–11). The protein complement of a monoparticle preparation consists of a discrete set of highly conserved polypeptides (1, 2, 9, 12, 13). An alternative approach to defining proteins closely associated with RNA is to chemically or photochemically crosslink them (14–17). The suite of RNP proteins identified by *in vivo* UV crosslinking displays remarkable overlap with the set detected by extraction of nuclei and gradient fractionation, although some species specific to each method of preparation are also seen (for reviews, see refs. 1 and 2).

Among the hnRNP proteins defined by the above biochemical criteria are the "core" ribonucleosome proteins. The widely used nomenclature of LeStourgeon *et al.* (9) designates prominent core hnRNP components from human cells in culture as A, B, and C proteins based on their migration pattern in sodium dodecyl sulfate (SDS)/polyacrylamide gels. It is thought that these proteins comprise a protein core of fixed stoichiometry with RNA wrapped around the outside of the particle (1, 2, 12, 18, 19). The type C proteins are especially tightly associated with RNA. They are, in comparison with the A and B proteins, the least sensitive to dissociation from RNA at increasing salt concentrations (1, 9) and are also among the most efficiently UV-crosslinked core components (16, 17, 20). C proteins are of particular biological interest because, in addition to their close physical association with hnRNA, they have recently been shown to be required for proper removal of intervening sequences in a mammalian *in vitro* splicing system (21).

The assembly and fate of the core hnRNP is of interest because it defines the first known posttranscriptional structure formed on the pathway leading to RNA maturation and transport out of the nucleus. Oocytes of the frog *Xenopus laevis* offer some special advantages for the study of RNA processing (22–26) and transport (27–30). We therefore initiated detailed molecular characterization of key proteins comprising the *Xenopus* nuclear hnRNP structure. In this report we describe the isolation and characterization of *Xenopus* C protein cDNAs and their RNA and protein products.[†]

MATERIALS AND METHODS

DNA Sequence. A *Xenopus* λgt10 ovary cDNA library (gift of D. Melton) (31) was screened with ^{32}P-labeled clone 4F4 human C protein cDNA insert (32). Hybridization criteria were 55°C in 5× SSPE (1× SSPE = 0.18 M NaCl/10 mM phosphate, pH 7.4/1 mM EDTA) followed by washing at 55°C with 2× SSPE. The sequence of pXEC1.3 was determined on both strands from overlapping clones by standard techniques (33–35).

DNA and RNA Analysis. DNA was prepared from the blood of a single female frog (36), and gel blots were performed on Pall Biodyne membranes with uniformly ^{32}P-labeled single-strand probes of $\approx 10^9$ dpm/μg. RNA was prepared from *Xenopus* liver, ovary, and A6 cultured cells essentially as described (37) with the addition of pelleting through a CsCl

Abbreviations: RNP, ribonucleoprotein; hnRNA, heterogeneous nuclear RNA; hnRNP, heterogeneous nuclear ribonucleoprotein; SDS, sodium dodecyl sulfate; nt, nucleotide(s).

*To whom reprint requests should be addressed.

[†]The sequence reported in this paper is being deposited in the EMBL/GenBank data base (IntelliGenetics, Mountain View, CA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03831).

```
GAATTCCGCCCCTCTAATCTCCGCGATAATCTAGCTCACATATTTATTTGAAGGTTAAAGGCCTCATAACAAAACGAATGGCAAGGGCTCATGTCTATCACACAGACGCTGTTGTGATTACTAGAAAGTTTCTGCAAGCGAAGAATTTTTTTTGAAACTTCA   162

                                 M  M  A  S  N  V  T  N  K  T  D  P  R  S  M  N  S  R  V  F  I  G  N  L  N  T  L  V  V  K  K  T  D  V  E  A  I  F  S  K  Y  G  K      43
ACTACCGCCAAAATTGCGTCCCCTCATTCCATCATGATGGCCAGTAATGTGACTAACAAGACGGATCCCCGTTCGATGAACTCGCGTGTATTTATTGGGAACCTTAATACGCTTGTTGTTAAGAAAACTGATGTAGAAGCAATCTTTTCAAAATATGGAAAG   324

I  V  G  C  S  V  H  K  G  F  A  F  V  Q  F  S  N  E  R  T  A  R  T  A  V  A  G  E  D  G  R  M  I  A  G  Q  V  L  D  I  N  L  A  A  E  P  K  A  N  R  S  K  T  G      97
ATTGTGGGCTGTTCTGTGCACAAGGGCTTTGCATTTGTGCAGTTTTCCAATGAACGCACTGCCCGTACAGCCGTTGCAGGTGAAGATGGGCGCATGATTGCAGGGCAAGTCCTGGATATCAATTTAGCTGCTGAACCTAAAGCAAACAGAAGCAAAACTGGT   486

V  K  R  S  A  A  D  M  Y  G  S  S  F  D  L  E  Y  D  F  P  R  D  Y  Y  D  S  Y  S  A  T  R  V  P  A  P  P  P  L  A  R  A  V  V  P  S  K  R  Q  R  V  S  G  N  A     151
GTCAAACGATCAGCTGCAGACATGTATGGGTCTTCCTTTGATTTGGAGTATGATTTCCCAAGAGATTACTATGACAGCTATTCTGCAACACGTGTACCAGCTCCTCCTCCATTAGCTCGGGCAGTAGTGCCATCAAAAAGGCAAAGAGTATCTGGAAATGCA   648

S  R  R  G  K  S  G  F  N  S  K  S  G  Q  R  G  G  S  S  K  S  S  R  L  K  G  D  D  L  Q  A  I  K  K  E  L  S  Q  I  K  Q  R  V  D  S  L  L  E  N  L  E  R  I  E     205
TCTCGGCGTGGTAAGAGTGGCTTTAACTCAAAAAGTGGCCAGCGAGGTGGTTCCTCAAAATCTAGTAGATTGAAGGGAGATGATCTTCAGGCAATCAAAAAGGAGCTCAGTCAGATAAAGCAGAGAGTAGATTCTCTCTTGGAAAACCTAGAAAGGATTGAG   810

R  D  Q  S  K  Q  D  T  K  L  D  D  D  Q  S  S  V  S  L  K  K  E  E  T  G  V  K  L  I  E  E  T  G  D  S  A  E  E  G  D  L  L  D  D  D  E  Q  G  E  D  T  L  E  E     259
CGTGACCAGTCAAAACAAGATACCAAATTAGATGATGACCAAAGCAGCGTTTCTTTAAAGAAAGAGGAGACTGGTGTTAAGCTGATAGAAGAAACAGGGGATTCTGCAGAGGAAGGAGACTTGCTTGATGATGATGAACAGGGTGAAGACACGCTTGAAGAA   972

I  K  D  G  D  K  E  T  E  E  G  E  D  E  G  D  S  A  N  E  E  D  S                                                                                                  282
ATTAAAGATGGAGACAAAGAAACAGAAGAGGGAGAAGATGAAGGAGACAGCGCTAACGAGGAAGACTCTTAAATTCATTAACCTTTCATGTAACTCTTCATCTGCTTGTCTTTCTGTCTTGTCTCATAGCACCTTTCTTAACAGTCCCTCAATCCATCCGCT  1134

GCTTTAAGCTTGTTTAAATATGCACCCTCCTATCCCTCAGCCTCCATTTCATTTTGATACCTGTTTGCGACTTCTAGAATAAAAGTGTATTGTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA                      1274
```

FIG. 1. Complete sequence of *Xenopus* C protein cDNA clone pXEC1.3 with translation of the major open reading frame.

cushion. Polyadenylylated RNA was selected by oligo(dT)-cellulose chromatography, fractionated in formaldehyde/agarose gels (38), transferred to filters, and hybridized as described for DNA. For primer extension 10 pmol of a primer complementary to nucleotides 231–245 was 5′-end-labeled to a specific activity of 10^9 dpm/μg using T4 polynucleotide kinase and extended with avian myeloblastosis virus reverse transcriptase at 42°C in 1 mM dNTPs with RNAsin at 0.5 units/μl. The reaction mixture was phenol extracted. Material excluded from G-50 Sephadex was analyzed by electrophoresis through a urea/polyacrylamide gel (36).

Translations. SP6 sense-strand RNA was transcribed from *Sal* I-linearized pXEC1.3 or pXEC1.0 (39). RNA was capped by inclusion of 2 mM bis(guanylyl)triphosphate (GpppG) in the transcription reaction with GTP at 500 μM and the other nucleotides at 2 mM. Purified RNA was suspended in water at 1 mg/ml. Twenty nanoliters of RNA per oocyte were injected, and they were incubated in OR2 buffer (82 mM NaCl/2.5 mM KCl/1 mM $MgCl_2$/1 mM $CaCl_2$/10 mM Hepes/1 mM NaH_2PO_4, pH 7.8) at 5°C for 30 min and then at 22°C for 8 hr. [^{35}S]Methionine was then added to a final concentration of 0.5–1 mCi/ml (1 Ci = 37 GBq), and the incubation was continued for 4–20 hr at room temperature. Oocyte nuclei and cytoplasms were prepared by manual dissection in J buffer (40) or by dissection after boiling for 1 min in OR2 buffer.

RESULTS AND DISCUSSION

Monoclonal antibodies specific for human C protein were previously used to identify and isolate a cDNA clone for human C protein (32). Here we used that 1.1-kilobase (kb) human C protein cDNA to identify homologous sequences in a library of cDNA clones prepared from *Xenopus* ovary polyadenylylated RNA (31). Cloned cDNAs of 1.0- and 1.3-kb lengths derived from two independent phage were subcloned, mapped, and sequenced. The 1274-nucleotide (nt) sequence of *Xenopus* C protein clone pXEC1.3 is shown in Fig. 1 together with the predicted amino acid sequence of the single long open reading frame. The open reading frame is preceded by a 195-nt 5′-untranslated segment that contains stop codons in all reading frames. The presence of these termination codons supports the assignment of Met-1 in Fig. 1 as the first possible initiation site for C protein synthesis. The methionine codon at position 2 in the open reading frame best conforms to the preferred consensus for translation initiation (41). The open reading frame codes for a protein of 30,916 Da, although the protein products produced by transcription and subsequent translation of the cloned sequence migrate in SDS gels with apparent molecular masses of 41 and 39 kDa (Fig. 3). The 3′-untranslated segment contains a single copy of the consensus signal for terminal cleavage and poly(A) addition, AATAAA (42), and a poly(A) tract that begins 20 nt downstream.

To confirm that the *Xenopus* clones isolated in this report are significantly similar to the human C probe, we determined the DNA sequence corresponding to the first 77 amino acid residues of the human clone. The DNA sequences display 83% identity and the protein sequences 92%, with all amino acid substitutions being functionally conservative (data not shown). While this work was in preparation the full human sequence was reported (43), and the sequence similarity extends throughout the protein coding sequence. The amino acid composition and expected isoelectric point of the *Xenopus* C protein deduced from our DNA sequence data agree remarkably well with those determined empirically for mammalian C proteins (13), whereas there is little resemblance to class A, B, or D RNP proteins. These data, taken together with immunological definition of the human clone and its high degree of similarity with the *Xenopus* open reading frame (17, 32), lead us to conclude that pXEC1.0 and 1.3 code for *Xenopus* hnRNP C protein.

C Protein Expression. The pXEC1.3 cDNA clone was used as a probe to identify C protein mRNA on RNA gel blots. A single size species of 1.3 ± 0.1 kb was detected in polyadenylylated RNA of A6 cells in culture (Fig. 2*A*), adult liver, or defolliculated oocytes (data not shown). The quantities of C protein RNA in stage V–VI oocytes and embryos at gastrula and neurula stages are all similar—about 1 × 10^6 transcripts per oocyte or embryo (R. Wagner, personal

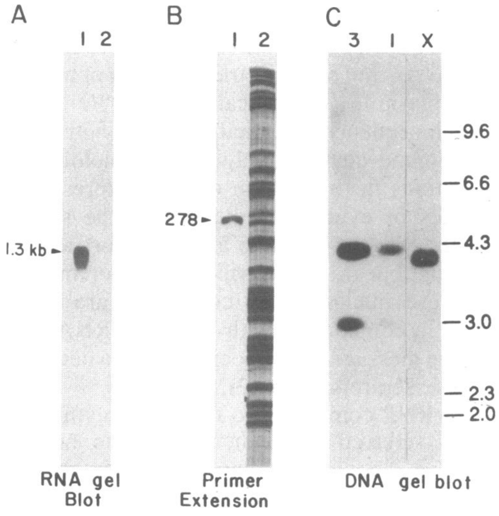

FIG. 2. (*A*) RNA gel blot of C protein RNA from the *Xenopus* A6 cell line. The size of the prominent RNA species hybridizing with cloned C protein cDNA probe is ≈1.3 kb based on the position of ribosomal RNA standards in adjacent lanes. Lane 1, Poly(A)$^+$ RNA fraction from 200 μg of total cellular RNA; lane 2, 25 μg of the poly(A)-depleted fraction of A6 RNA. (*B*) Primer-extension using C protein RNA as template. A ^{32}P-end-labeled C protein primer was hybridized with 25 μg of total RNA from *Xenopus* liver, the primer was extended by reverse transcriptase, and extension products were displayed on a sequencing gel calibrated by a DNA sequence ladder generated by extension of the same primer used in RNA analysis. Lane 1, extension products; lane 2, G sequence ladder. (*C*) DNA gel blot. The lanes labeled 1 and 3 are reconstruction standards containing the equivalent of one and three copies of C protein cloned DNA per haploid genome for a 5-μg sample of *Xenopus* nuclear DNA. Lane X contains 5 μg of *Xenopus* nuclear DNA digested with *Xba* I.

communication). This prevalence is typical of the majority of maternal poly(A)$^+$ RNAs in *Xenopus* stage-VI oocytes (44).

The 5′ end of C protein mRNA has a long untranslated leader that is represented in clone pXEC1.3, whereas pXEC1.0 possesses a shorter 5′-untranslated segment. To better define the 5′ end of this RNA, a labeled primer was hybridized with total liver RNA and extended with avian myeloblastosis virus reverse transcriptase (Fig. 2*B*). A prominent product 33 nt longer than pXEC1.3 was seen; this is consistent with the mRNA length seen in the RNA gel blot experiments and indicates that the 5′-untranslated segment is ≈228 nt in length. This size constitutes a minimum length, and definitive assignment of the transcription start will require cloning and transcript mapping of genomic sequences.

One RNA Species Produces Two C Proteins. RNP preparations from mammalian cells contain two distinct C protein species that can be resolved in SDS/polyacrylamide gels, and these species occur in a relatively reproducible ratio of ≈1:3, with the more slowly migrating species present in lower amounts. We investigated the synthesis of *Xenopus* C protein from defined cloned mRNAs transcribed *in vitro* by SP6 polymerase. These were translated in rabbit reticulocyte extracts and in *Xenopus* oocytes. RNA transcribed from pXEC1.3 contains the long 195-nt 5′-untranslated segment, and in both translation systems this RNA produced two polypeptide species that migrate with apparent molecular masses of 41 and 39 kDa in a ratio similar to *in vivo* RNP preparations from mammalian cells (Fig. 3 *A* and *B*). Surprisingly, only the 39-kDa product is produced when the shorter pXEC1.0 RNA is the template (Fig. 3*B*). As the two RNAs differ only in the lengths of their 5′-untranslated sequences, it seems unlikely that the protein species arise from some posttranslational modification. Although it is formally possible that the 41-kDa product could initiate from an upstream methionine shifted into frame by an SP6 transcription error, a more attractive alternative is that the two protein species are produced by the use of more than one initiation codon in the major open reading frame (methionine 1 or 2 versus methionine 15, for example). If this explanation is correct, it suggests that the mechanism by which the initiation choice is made must depend, at least in part, on the 5′-untranslated leader. Furthermore, the choice is made similarly, both quantitatively and qualitatively, in frog oocytes and rabbit reticulocyte lysates. This choice is probably not a peculiarity of the *Xenopus* gene because C protein RNA prepared by hybrid selection from cultured human cells was recently shown to produce both C1 and C2 in a reticulocyte translation system (43). The nature of the difference between C1 and C2 proteins and the molecular mechanism by which it is generated should be decisively resolved by direct protein sequence determinations together with site-directed mutagenesis of 5′-untranslated sequences and proposed initiation codons.

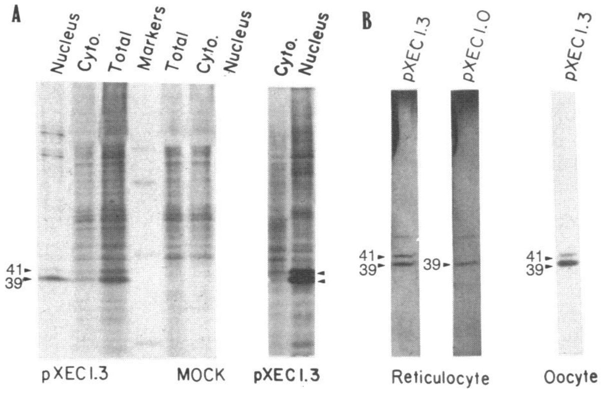

FIG. 3. (*A*) *In vivo* translation and subcellular localization proteins produced from SP6-transcribed C protein mRNA. Individual oocytes were dissected as indicated and displayed on an SDS/polyacrylamide gel. Nuclear and cytoplasmic fractions from a second independent experiment are shown at right. (*B*) *In vitro* translation of two species of C protein SP6 RNAs generated from clones with differing 5′-untranslated segments. RNAs containing either pXEC1.0 or pXEC1.3 cDNA inserts were produced using SP6 polymerase *in vitro* and then translated in a rabbit reticulocyte lysate containing [^{35}S]methionine. The labeled products are displayed on an SDS/polyacrylamide gel as indicated. pXEC1.0 contains a truncated 5′-untranslated region beginning at nucleotide 141 of the sequence in Fig. 1.

Nuclear Localization. In mammalian somatic cells RNP proteins are localized in the nuclei of interphase cells. To examine the subcellular localization of C protein in *Xenopus* oocytes, the cDNA sequences were inserted into the bacteriophage RNA polymerase vector pSP65. Capped RNA was transcribed by SP6 polymerase, and the RNA was microinjected into the cytoplasm of stage-VI *Xenopus* oocytes. Newly synthesized proteins were labeled by incubating the oocytes in [^{35}S]methionine. The distribution of labeled C protein was measured by dissecting oocytes into cytoplasmic and nuclear fractions and subjecting the samples to SDS gel electrophoresis (Fig. 3*A*). In most experiments 80–95% of both newly synthesized C protein species were in the germinal vesicle fraction after 4–20 hr of labeling. The nucleus composes ≈12% of the oocyte volume (excluding yolk platelets), so the observed distribution of C protein corresponds to a 7- to 18-fold concentration in the nuclear compartment. Other newly synthesized proteins encoded by endogenous and heterologous RNAs for nonnuclear proteins are not localized in the germinal vesicle (Fig. 3 and B.W., unpublished data). The data presented here do not provide a direct measure of the kinetics of import into the nucleus, but the shortest labeling periods showed accumulation of >85% of both C protein species within 4 hr, suggesting that import is quite rapid.

A characteristic of some rapidly accumulated nuclear proteins is the presence of one or more copies of a small nuclear localization sequence (for review see ref. 45). A search of the *Xenopus* C protein revealed one candidate sequence beginning at residue 141, Pro-Ser-Lys-Arg-Gln-Arg-Val, that is a 5/7 match with the consensus simian virus 40 large tumor antigen nuclear localization tag (46) and also resembles related sequences in the nuclear localization region of *Xenopus* nucleoplasmin (47). Although this sequence appears to be a good candidate for the nuclear target sequence (differences from tested sites are conservative or appear at the most flexible residue), its physiological significance remains to be determined.

C Protein Sequences in the Genome. Mammalian genomes contain numerous sequences that hybridize with the human C protein clone (32), and by analogy with genomic cloning data for other human RNPs, it seems likely that many C protein sequences will turn out to be pseudogenes (C. Morandi, personal communication). In contrast to the mammalian data, gel blots of DNA from a single frog show a simple pattern of one-to-three fragments with different restriction enzyme digests. Comparison with known copy-number standards (Fig. 2*C*) and the results of screening frog genomic libraries all indicate that the *Xenopus* genome contains a very small number of C protein genes—probably only two or three per haploid genome. Reducing the hybridization and filter-washing stringencies by 10°C failed to identify any new *Xenopus* DNA fragments that would indicate divergent but related genes.

C Protein Structure. Several features of the *Xenopus* C protein sequence are of interest with respect to its interaction with RNA, its possible contributions to RNP assembly and architecture, and its localization in the nucleus. Three potential domains can be identified, and these domains are noted in Fig. 4. The amino-terminal domain of ≈90 residues shares with other RNA-binding proteins a similar array of aromatic residues as well as a specific octapeptide sequence that has been found in virtually all nuclear RNA-binding proteins for which sequence information is available. The octapeptide was first noted in mammalian A protein sequence (49) and has been postulated to mediate binding of yeast poly(A) binding protein with poly(A) (50). The occurrence of this hallmark feature in several different RNP proteins from phylogenetically diverse organisms is summarized in Table 1. The consensus octapeptide includes three aromatic residues that might be expected to interact with RNA by intercalating between bases, whereas nearby positively charged residues would be expected to facilitate binding by countering negative charges on the nucleic acid phosphate backbone. RNA binding has not yet been directly demonstrated for the amino-terminal domain of C protein, but similar 90- to 100-residue domains of other RNP proteins have been shown to bind single-strand nucleic acids *in vitro* (2, 62).

The presence of the octapeptide sequence in diverse RNA-binding proteins together with the potential nucleic acid-binding character of the sequence itself, suggest that the sequence plays a role in RNA–protein interactions. However, the motif of aromatic amino acids with a nearby basic residue is not confined to the conserved sequence and also appears at several other points (arrows in Fig. 4, domain I) within the larger 90–100-amino acid "RNA domains" of each of these RNP proteins. The positions of the additional basic-aromatic elements are similar in other RNP proteins. This raises the possibility that the RNA-binding character of these domains is distributed among the several basic-aromatic elements, rather than being localized within the octapeptide region. In support of this possibility, the first mutational studies of yeast poly(A)-binding protein show that neither mutation of conserved aromatic residues within the octapeptide nor deletion of the entire octapeptide eliminates RNA binding *in vitro* or the capacity to complement a poly(A)-binding protein-deficient strain *in vivo* (62).

A distinction between the proposed RNA-binding domain of C protein and that of all other such domains identified to date is its net charge. C protein is acidic overall, but this is entirely due to the concentration of aspartic and glutamic acid residues in the hydrophilic carboxyl-terminal domain. The amino-terminal RNA domain alone is, in fact, highly basic with basic-to-acidic content of 12:4. It seems possible that this local positive charge facilitates RNA binding, and this charge may contribute to the comparatively tight RNA binding that characterizes C proteins relative to A and B proteins (9). The A protein RNA domains, for example, are much less basic with basic-to-acidic content of 13:13 and 11:9 (55). The proposed middle domain is distinctive among RNA-binding proteins characterized to date. The middle domain is basic but lacks an RNP consensus.

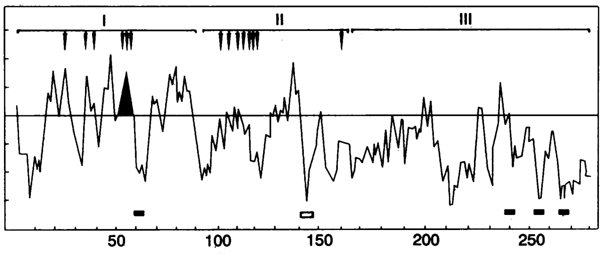

FIG. 4. C protein hydropathy plot and sequence features. The hydropathy values of Kyte and Doolittle (48) were used with a window of seven. The protein domains I–III discussed in the text are indicated. Vertical arrows mark aromatic residues. ■, Potential type II casein kinase phosphorylation sites; □, potential nuclear localization tag. The proposed RNP consensus octamer is indicated by dark shading of the hydropathy profile.

Table 1. RNA-binding protein homology

Protein	Organism (ref.)	Amino acid	Sequence
Poly(A)-binding protein	Yeast (50, 51)	78–93	SLGYAVNFNDHEAGRK
		165–180	SKGFGFVHFEEEGAAK
		258–273	LKGFGFVNYEKHEDAV
		361–376	SKGFGFVCFSTPEEAT
	Human (52)	51–66	SLGYAYVNFQQPADAE
		137–152	SKGYGFVHFETQEAAE
		227–242	SKGFGFVSFERHEDAQ
		329–344	SKGFGFVCFSSPEEAT
SSB-1	Yeast (53)	234–249	NRGMAFVTFSGENVDI
Nucleolin	Hamster (54)	346–361	NRKFGYVDFESAEDLE
		428–443	SKGIAYIEFKSEADAE
		521–536	SKGYAFIEFASFEDAK
		608–623	SKGFGFVAFNSEEDAK
UP1	Calf (49)	53–68	SRGFGFVTYATVEEVD
		144–159	KRGFAFVTFDDHDSVD
A1/hnRNP	Rat and human (55, 56)	54–69	SRGFGFVTYATVEEVD
		145–160	KRGFAFVTFDDHDSVD
UP2/hnRNP?	Human (57)	212–225	RRGFCFITFNQEEP
C/hnRNP	Human (43)	49–64	HKGFAFVQYVNERNAR
	Frog (Fig. 1)	50–65	HKGFAFVQFSNERTAR
U1 snRNP 70 kDa	Human (58)	320–335	PRGYAFIEYEHERDMH
U2 snRNPB″	Human (59)	48–63	MRGQAFVIFKELSSTN
U1 snRNPA	Human (60)	51–66	MRGQAFVIFKEVSSAT
P9	Fly (61)	71–86	SRGFGFITYSHSSMID
		163–178	KRGFAFVEFDDYDPVD
Core consensus			KGFGFVXF
Frequent changes			R YAYI Y

The carboxyl-terminal domain of C protein is strikingly different from the first two and is responsible for the acidic character of the protein. This conforms to the general organization of other nuclear RNA-binding proteins in which the putative ancestral RNA-binding domain(s) are at the amino terminus, whereas distinctive, protein-specific domains are located at the carboxyl terminus. The proposed third domain of C protein is ≈100 amino acids in length and contains 39 acidic residues, 13 basic residues, and no aromatic amino acids. Moreover, three of four potential sites for phosphorylation by a casein type II kinase activity (63) are located in the third domain at Ser-240, Thr-256, and Thr-267. It is known that mammalian C proteins are phosphorylated *in vivo* and are substrates *in vitro* for a type II casein kinase (64, 65). These posttranslational modifications would further increase the negative charge in the carboxyl-terminal domain, and they present the obvious possibility of regulating function.

The carboxyl-terminal domains of the RNPs have been postulated to mediate protein–protein interactions (for review, see ref. 2). The C protein carboxyl-terminal domain is hydrophilic and acidic. Although these properties contrast sharply with the analogous domains of other core RNP proteins, such properties are reminiscent of another *Xenopus* nuclear protein, nucleoplasmin (47). Nucleoplasmin facilitates nucleosome assembly (66, 67), apparently by neutralizing positive charge on core histones. This structural similarity between nucleoplasmin and C protein makes it interesting to consider a potential functional analogy in which C protein may play a role in ribonucleosome or spliceosome assembly by neutralizing basic proteins such as the A and B core hnRNPs. In this case, however, C protein would remain as a component of the ribonucleosome, whereas nucleoplasmin does not remain as part of the assembled nucleosome.

The availability of cloned cDNAs coding for several nuclear RNA-binding proteins [poly(A)-binding protein of yeast (50, 51) and *Xenopus* (R. Moon, unpublished work), A protein of rat (55) and *Xenopus* (D. Ruff and B.W., unpublished data), and C protein of human (43) and *Xenopus*] should now permit decisive tests for the function of different RNP domains and of explicit features within a domain.

This work was supported by a Postgraduate Fellowship from the National Research Council of Canada (National Sciences and Engineering Research Council) and the Gordon Ross Fund to F.P. and grants from the National Institutes of Health and Rita Allen Foundation to B.W.

1. Dreyfuss, G. (1986) *Annu. Rev. Cell Biol.* **2**, 459–498.
2. Chung, S. Y. & Wooley, J. (1986) *Proteins Struct. Funct. Genet.* **1**, 195–210.
3. McKnight, S. L. & Miller, O. L., Jr. (1976) *Cell* **8**, 305–319.
4. Malcolm, D. B. & Sommerville, J. (1977) *J. Cell. Sci.* **24**, 143–165.
5. Beyer, A. L., Bouton, A. H. & Miller, O. L., Jr. (1981) *Cell* **26**, 155–165.
6. Foe, V. E., Wilkinson, L. E. & Laird, C. D. (1976) *Cell* **9**, 131–146.
7. Sommerville, J. (1981) in *The Cell Nucleus*, ed. H. Busch (Academic, New York), pp. 1–57.
8. Samarina, O. P., Molnar, J., Lukanidin, E. M., Bruskov, V. I., Krichevskaya, A. A. & Georgiev, G. P. (1967) *J. Mol. Biol.* **27**, 187–191.
9. Beyer, A. L., Christensen, M. E., Walker, B. W. & LeStourgeon, W. M. (1977) *Cell* **11**, 127–138.
10. Pederson, T. (1974) *J. Mol. Biol.* **83**, 163–183.
11. Martin, T., Billings, P., Pullman, J., Stevens, B. & Kinniburgh, A. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 899–909.
12. Lothstein, L., Arenstorf, H. P., Chung, S. Y., Walker, B. W., Wooley, J. C. & LeStourgeon, W. M. (1985) *J. Cell. Biol.* **100**, 1570–1581.
13. Wilk, H., Werr, H., Friedrich, D., Kiltz, H. H. & Schafer, K. P. (1985) *Eur. J. Biochem.* **146**, 71–81.
14. Setyono, B. & Greenberg, J. R. (1981) *Cell* **24**, 775–783.
15. Mayrand, S. & Pederson, T. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2208–2212.
16. Van Eekelen, C. A. G., Mariman, E. C. M., Reinders, R. J. & Van Venrooij, W. J. (1981) *Eur. J. Biochem.* **119**, 461–467.
17. Dreyfuss, G., Choi, Y. D. & Adam, S. A. (1984) *Mol. Cell. Biol.* **4**, 1104–1114.
18. LeStourgeon, W. M., Lothstein, L., Walker, B. W. & Beyer, A. L. (1981) in *Nuclear Particles, Part B*, ed. Busch, H. (Academic, New York), pp. 49–87.
19. Choi, Y. D. & Dreyfuss, G. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 7471–7475.
20. Van Eekelen, C. A., Rieman, T. & van Venrooij, W. J. (1981) *FEBS Lett.* **130**, 223–226.
21. Choi, Y. D., Grabowski, P. J., Sharp, P. A. & Dreyfuss, G. (1986) *Science* **231**, 1534–1539.
22. Green, M. R., Maniatis, T. & Melton, D. A. (1983) *Cell* **32**, 681–694.
23. Krieg, P. A. & Melton, D. A. (1984) *Nature (London)* **308**, 203–206.
24. Galli, G., Hofstetter, H., Stunnenberg, H. G. & Birnstiel, M. L. (1983) *Cell* **34**, 823–828.
25. Conway, L. & Wickens, M. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 3949–3953.
26. Adeniyi-Jones, S. & Zasloff, M. (1985) *Nature (London)* **317**, 81–84.
27. Zasloff, M., Rosenberg, M. & Santos, T. (1982) *Nature (London)* **300**, 81–84.
28. Neuman de Vegvar, H. E., Lund, E. & Dahlberg, J. E. (1986) *Cell* **47**, 259–266.
29. Georgiev, O., Mous, J. & Birnstiel, M. L. (1984) *Nucleic Acids Res.* **12**, 8539–8551.
30. Zasloff, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6436–6440.
31. Rebagliati, M. R., Weeks, D. L., Harvey, R. P. & Melton, D. A. (1985) *Cell* **42**, 769–777.
32. Nakagawa, T. K., Swanson, M. S., Wold, B. J. & Dreyfuss, G. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2007–2011.
33. Messing, J (1983) *Methods Enzymol.* **101**, 20–78.
34. Hattori, M. & Sakaki, Y. (1986) *Anal. Biochem.* **152**, 232–238.
35. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
36. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
37. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299.
38. Lehrach, H., Diamond, D., Wozney, J. M. & Boedtker, H. (1977) *Biochemistry* **16**, 4743–4751.
39. Melton, D. A., Krieg, P. A., Rebagliati, M. R., Maniatis, T., Zinn, K. & Green, M. R. (1984) *Nucleic Acids Res.* **12**, 7035–7056.
40. DeRobertis, E. M., Black, P. & Nishikura, K. (1981) *Cell* **23**, 89–93.
41. Kozak, M. (1983) *Microbiol. Rev.* **47**, 1–45.
42. Fitzgerald, M. & Shenk, T. (1981) *Cell* **24**, 251–260.
43. Swanson, M. S., Nakagawa, T., LeVan, K. & Dreyfuss, G. (1987) *Mol. Cell. Biol.* **7**, 1731–1739.
44. Davidson, E. H. (1987) *Gene Activity in Early Development* (Academic, New York).
45. Dingwall, C. & Laskey, R. (1986) *Annu. Rev. Cell Biol.* **2**, 365–388.
46. Smith, A. E., Kalderon, D., Roberts, B. L., Colledge, W. H., Edge, M., Gillett, P., Markham, A., Paucha, E. & Richardson, W. D. (1985) *Proc. R. Soc. London Ser. B* **221**, 43–58.
47. Dingwall, C., Dilworth, S. M., Black, S. J., Kearsey, S. E., Cox, L. S. & Laskey, R. A. (1987) *EMBO J.* **6**, 69–74.
48. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
49. Merrill, B. M., Lopresti, M. B., Stone, K. L. & Williams, K. R. (1986) *J. Biol. Chem.* **261**, 878–883.
50. Adam, S. A., Nakagawa, T. Y., Swanson, M. S. & Dreyfuss, G. (1986) *Mol. Cell. Biol.* **6**, 2932–2943.
51. Sachs, A. B., Bond, M. W. & Kornberg, R. D. (1986) *Cell* **45**, 827–835.
52. Grange, T., Martins de Sa, C., Oddos, J. & Pictet, R. (1987) *Nucleic Acids Res.* **15**, 4771–4787.
53. Jong, A. Y.-S., Clark, M. W., Gilbert, M., Oehm, A. & Campbell, J. L. (1987) *Mol. Cell. Biol.* **7**, 2947–2955.
54. LePeyre, B., Bourbon, H. & Amalric, F. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1472–1476.
55. Cobianchi, F., SenGupta, D. N., Zmudzka, B. Z. & Wilson, S. H. (1986) *J. Biol. Chem.* **261**, 3536–3543.
56. Buvoli, M., Biamonti, G., Tsoulfas, P., Bassi, M. T., Ghetti, A., Riva, S. & Morandi, C. *Nucleic Acids Res.* (1988) **16**, 3751–3770.
57. Lahiri, D. K. & Thomas, J. O. (1986) *Nucleic Acids Res.* **14**, 4077–4094.
58. Thiessen, H., Etzerodt, M., Reuter, R., Schneider, C., Lottspeich, F., Argos, P., Luehrmann, R. & Philipson, L. (1986) *EMBO J.* **5**, 3209–3217.
59. Habets, W. J., Sillekens, P. T. G., Hoet, M. H., Schalken, J. A., Roebroek, A. J. M., Leunissen, J. A. M., Van de Ven, W. J. M. & van Venrooij, W. J. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2421–2425.
60. Sillekens, P. T. G., Habets, W. J., Beijer, R. P. & van Venrooij, W. J. (1987) *EMBO J.* **6**, 3841–3848.
61. Haynes, S. R., Rebbert, M. L., Mozer, B. A., Forguignon, F. & Dawid, I. B. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1819–1823.
62. Sachs, A. B., Davis, R. W. & Kornberg, R. D. (1987) *Mol. Cell. Biol.* **7**, 3268–3276.
63. Hathaway, G. M. & Traugh, J. A. (1982) *Curr. Top. Cell. Regul.* **21**, 101–127.
64. Dreyfuss, G., Choi, Y. D. & Adam, S. A. (1984) *Mol. Cell. Biol.* **4**, 1104–1114.
65. Holcomb, E. R. & Friedman, D. L. (1984) *J. Biol. Chem.* **259**, 31–40.
66. Earnshaw, W. C., Honda, B. M. & Laskey, R. A. (1980) *Cell* **21**, 373–383.
67. Kleinschmidt, J. A., Fortkamp, E., Krohne, G., Zentgraf, H. & Franke, W. W. (1985) *J. Biol. Chem.* **260**, 1166–1176.