MOLECULAR AND CELLULAR BIOLOGY, May 1987, p. 1731–1739
0270-7306/87/051731-09$02.00/0

Vol. 7, No. 5

Primary Structure of Human Nuclear Ribonucleoprotein Particle C Proteins: Conservation of Sequence and Domain Structures in Heterogeneous Nuclear RNA, mRNA, and Pre-rRNA-Binding Proteins

MAURICE S. SWANSON, TERRY Y. NAKAGAWA, KAY LEVAN, AND GIDEON DREYFUSS*

Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, Evanston, Illinois 60201

Received 10 October 1986/Accepted 5 February 1987

In the eucaryotic nucleus, heterogeneous nuclear RNAs exist in a complex with a specific set of proteins to form heterogeneous nuclear ribonucleoprotein particles (hnRNPs). The C proteins, C1 and C2, are major constituents of hnRNPs and appear to play a role in RNA splicing as suggested by antibody inhibition and immunodepletion experiments. With the use of a previously described partial cDNA clone as a hybridization probe, full-length cDNAs for the human C proteins were isolated. All of the cDNAs isolated hybridized to two poly(A)$^+$ RNAs of 1.9 and 1.4 kilobases (kb). DNA sequencing of a cDNA clone for the 1.9-kb mRNA (pHC12) revealed a single open reading frame of 290 amino acids coding for a protein of 31,931 daltons and two polyadenylation signals, AAUAAA, ~400 base pairs apart in the 3′ untranslated region of the mRNA. DNA sequencing of a clone corresponding to the 1.4-kb mRNA (pHC5) indicated that the sequence of this mRNA is identical to that of the 1.9-kb mRNA up to the first polyadenylation signal which it uses. Both mRNAs therefore have the same coding capacity and are probably transcribed from a single gene. Translation in vitro of the 1.9-kb mRNA selected by hybridization with a 3′-end subfragment of pHC12 demonstrated that it by itself can direct the synthesis of both C1 and C2. The difference between the C1 and C2 proteins which results in their electrophoretic separation is not known, but most likely one of them is generated from the other posttranslationally. Since several hnRNP proteins appeared by sodium dodecyl sulfate-polyacrylamide gel electrophoresis as multiple antigenically related polypeptides, this raises the possibility that some of these other groups of hnRNP proteins are also each produced from a single mRNA. The predicted amino acid sequence of the protein indicates that it is composed of two distinct domains: an amino terminus that contains what we have recently described as an RNP consensus sequence, which is the putative RNA-binding site, and a carboxy terminus that is very negatively charged, contains no aromatic amino acids or prolines, and contains a putative nucleoside triphosphate-binding fold, as well as a phosphorylation site for casein kinase type II. The RNP consensus sequence was also found in the yeast poly(A)-binding protein (PABP), the heterogeneous nuclear RNA-binding proteins A1 and A2, and the pre-rRNA binding protein C23. All of these proteins are also composed of at least two distinct domains: an amino terminus, which possesses one or more RNP consensus sequences, and a carboxy terminus, which is unique to each protein, being very acidic in the C proteins and rich in glycine in A1, A2, and C23 and rich in proline in the poly(A)-binding protein. These findings suggest that the amino terminus of these proteins possesses a highly conserved RNA-binding domain, whereas the carboxy terminus contains a region essential to the unique function and interactions of each of the RNA-binding proteins.

Heterogeneous nuclear RNAs (hnRNAs) become associated with a set of abundant nuclear proteins as they are being transcribed by RNA polymerase II. These RNA-protein complexes, referred to as heterogeneous nuclear ribonucleoprotein particles (hnRNPs), were originally characterized by electron microscopy and sucrose gradient sedimentation (10, 15, 30, 32). hnRNPs viewed in the electron microscope are composed of linear arrays of globular protein units, about 20 nm in diameter, which are connected by RNase-sensitive strands (14, 23). The hnRNA in nucleoplasm has been found to sediment in sucrose gradients in 50 to 100 mM NaCl as a heterodispersed material between 30 and 250S (2, 36, 42). Mild RNase digestion converts this dispersed fast-sedimenting hnRNA to more discrete particles (monoparticles) that sediment at ~30S. Biochemical analyses indicated that monoparticles are composed of hnRNA fragments and proteins, including a distinct nonchromatin group of proteins in the 30- to 43-kilodalton (kDa) range (2, 3, 13, 19, 27, 28). As methods for the electrophoretic resolution of these proteins became more refined, the number of individual polypeptides which make up the 30S particle increased (10, 46). Two-dimensional gel electrophoresis has suggested that the 30- to 43-kDa proteins are modified by charge and partly by phosphorylation (46). However, because of the incomplete purification of the particles, uncertainties about their true composition remained.

Recently, we have used monoclonal antibodies against several hnRNP proteins to immunopurify and characterize the hnRNP complex from vertebrate cell nuclei (4, 10). This confirmed and extended some of the earlier observations and provided a clearer definition of the hnRNP complex. Of particular interest are the two immunologically related hnRNP proteins, C1 and C2 (in human cells, 41,000 and

* Corresponding author.

43,000 daltons, respectively), that are tightly associated with RNA (2, 5, 12). The C proteins appear to be associated with it during splicing, since both pre-mRNA, spliced mRNA, and splicing intermediates can be immunoprecipitated from nucleoplasm with anti-C protein monoclonal antibodies (Y. D. Choi and G. Dreyfuss, unpublished results). In fact, recent observations suggest an important role for these proteins in RNA splicing; a monoclonal antibody to the C proteins inhibits in vitro splicing of an mRNA precursor, and depletion of these proteins from the splicing extract abolishes its capacity to splice pre-mRNA (6).

We have previously described the isolation of a partial C protein clone, pHC4F4, and the identification of two hybridizing poly(A)$^+$ RNAs of 1.9 and 1.4 kilobases (kb) (33). In the present work, the pHC4F4 clone was used as a hybridization probe for the isolation of cDNA clones containing the entire coding region. We report here the complete primary structure of the C proteins and describe experiments which indicate that a single coding sequence produces both C1 and C2. The sequence reveals a protein with two distinct domains: an amino terminus containing the highly conserved RNP consensus sequence, which most likely constitutes part of the RNA-binding site (1), and an acidic and hydrophilic carboxy terminus, which contains a putative nucleoside triphosphate (NTP)-binding fold and a protein kinase phosphorylation site. We also report the finding of an RNP consensus sequence in C23, a major 110 kDa nucleolar pre-rRNA-binding protein.

MATERIALS AND METHODS

Cell culture and labeling. HeLa S3 (human) cells were cultured in monolayer to subconfluent densities as previously described (4, 12). Cells were labeled with [^{35}S]methionine at 20 μCi/ml for 20 h in Dulbecco modified Eagle medium containing 1/10 the normal methionine level and 5% fetal calf serum.

Cell fractionation and immunoprecipitation. For the preparation of labeled C proteins from cells, the nucleoplasmic fraction was isolated as previously described (4), and immunoprecipitations were performed with 1% Empigen BB–1 mM EDTA–0.1 mM dithiothreitol in phosphate-buffered saline (5). The preparation and characterization of the 4F4 monoclonal antibody to the C proteins have also been previously described (5, 12).

Isolation of cDNA clones. The cDNA library used for the isolation of full-length or nearly full-length cDNA clones (HC5 and HC12, described below) was prepared from poly(A)$^+$ RNA isolated from the human premonocytic cell line U937. The library was constructed and kindly provided by Neal Farber of Biogen Inc., Cambridge, Mass. The library was screened with the nick-translated (39) pHC4F4 insert (33) by hybridization at 42°C, in 50% formamide–5× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate, pH 7.0)–0.2% sodium dodecyl sulfate (SDS)–5× Denhardt solution (9) containing calf thymus DNA (100 μg/ml) and poly(A) (1 μg/ml). Blots were washed in 1× SSC–0.1% SDS at 65°C. Approximately 100 positive clones were picked from an initial screen of 5×10^5 plaques, and 20 of these were plaque purified and characterized further by restriction mapping. All of these clones contained a 475-base-pair (bp) *Hin*fI fragment which hybridized to pHC4F4 (nucleotides 185 to 660; see Fig. 4) and were found to be overlapping sequences. The clones containing the largest insert from each class (see Results), HC5 and HC12, were chosen for subcloning to create pHC5 and pHC12.

RNA blot hybridizations. Poly(A)$^+$ RNA was prepared as previously described (33), resolved by electrophoresis on 1.4% agarose gels in the presence of formaldehyde (25), and blotted onto nitrocellulose (26). Hybridization probes were prepared by nick translation with [^{32}P]dCTP (39) of various fragments of the pHC12 insert. Hybridization and washing were performed as previously described (1).

Hybrid selection, in vitro translation, and gel electrophoresis. Poly(A)$^+$ RNA was prepared from HeLa or U937 cytoplasm by phenol extraction in the presence of 10 mM vanadyl-adenosine and chromatography on oligo(dT)-cellulose (33). Hybridization selection was performed by a modification of the technique of Ricciardi et al. (38) as previously described (33). The pHC12 fragments used for hybrid selection were the entire insert (1,686 bp), the 5′-end-specific fragment from *Eco*RI-*Bal*I (120 bp), or the 3′-end fragment from *Nsi*I-*Eco*RI (150 bp); the last two fragments were ligated into concatamers with T4 DNA ligase (26) before filter binding to increase the efficiency of retention on nitrocellulose. Optimized in vitro translations with micrococcal nuclease-treated rabbit reticulocyte lysate (37) were performed for 60 min at 37°C in 25-μl reaction volumes, whereas those with wheat germ were for 60 min at 28°C in 100-μl reaction volumes. Proteins were analyzed by SDS-polyacrylamide (12.5%) gel electrophoresis (SDS-PAGE), and fluorography of ^{35}S-labeled material was performed as previously described (11).

DNA sequence analysis. The overlapping cDNA clones, HC4F4, HC5, and HC12, were subcloned into pGEM1 as *Eco*RI inserts, and smaller overlapping DNA fragments were cloned into M13mp18 and M13mp19 (31). Sequencing reactions were carried out by the dideoxy chain termination method (43) with both the Klenow fragment of DNA polymerase I and avian myeloblastosis virus reverse transcriptase. Primer hybridization, polymerase reactions, and the dATP chase were all performed at 50°C for both types of enzyme. Sequence analysis was performed with the University of Wisconsin Genetics Computer Group Sequence Analysis Programs, and secondary structure predictions were done with the ChouFas and ChouPlot programs run on a model 7550A graphics plotter (Hewlett-Packard Co., Palo Alto, Calif.).

RESULTS

Production of C1 and C2 from the same mRNA. We have previously described a partial cDNA clone, pHC4F4, which hybrid selected poly(A)$^+$ RNA that translated in vitro both C1 and C2 (33). The cDNA, 1.1 kb in size, hybridized on a blot to two poly(A)$^+$ RNAs of 1.9 and 1.4 kb. With pHC4F4 as a hybridization probe, a 1.6-kb cDNA clone, termed HC12, was isolated by screening a λgt10 human cDNA library prepared from the premonocytic cell line U937. Restriction analysis of pHC12 was carried out, and different fragments were used for RNA blot analysis of HeLa or U937 poly(A)$^+$ RNA. Whereas the whole HC12 and a fragment containing only 120 bp of the 5′ end hybridized to both 1.9- and 1.4-kb mRNAs (Fig. 1A and B), a fragment containing the 150 bp of the 3′ end of the clone (*Nsi*I to *Eco*RI; see Fig. 3) hybridized only to the 1.9-kb mRNA (Fig. 1C). To identify the protein encoded by the 1.9-kb mRNA, the *Nsi*I-*Eco*RI fragment was used to hybrid select this mRNA, and the in vitro translation products produced in a rabbit reticulocyte lysate were analyzed by SDS-PAGE (Fig. 2). Interestingly, the mRNA selected by the *Nsi*I-*Eco*RI fragment, which hybridizes only to the 1.9-kb mRNA, produced both the C1

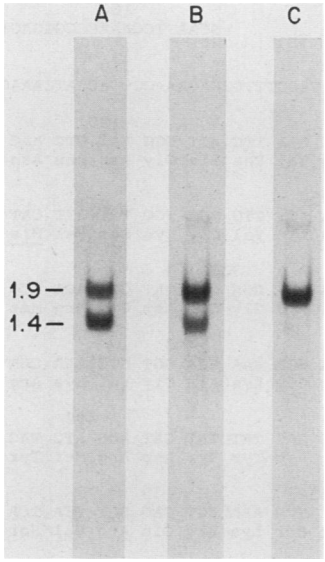

FIG. 1. RNA blot analysis of pHC12. Poly(A)$^+$ RNA from HeLa cells was resolved on a 1.4% formaldehyde agarose gel and blotted onto nitrocellulose. The blot was probed with the nick-translated *BalI-EcoRI* large fragment containing most of the coding sequence (A) (see Fig. 3), *EcoRI-BalI* 5′-end noncoding fragment (B) (120 bp), and *NsiI-EcoRI* 3′-end noncoding fragment (C) (150 bp). The sizes of the hybridizing RNAs are estimated by using *HindIII*-digested λDNA as a marker.

and C2 proteins (Fig. 2C). Moreover, the ratio of products was similar to that found in vivo (Fig. 2A) and by translation in vitro of total unselected mRNA (Fig. 2B). These findings suggest that C1 and C2 are produced from the same 1.9-kb mRNA either by differential use of the coding information in

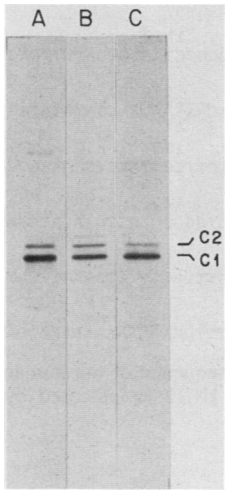

FIG. 2. Hybrid selection and in vitro translation with pHC12. HeLa poly(A)$^+$ RNA was prepared and hybridized to filter-bound DNAs and translated in vitro in a rabbit reticulocyte lysate after elution. The ^{35}S-labeled polypeptides were immunoprecipitated with the anti-C protein monoclonal antibody, 4F4. (A) C proteins immunoprecipitated directly from HeLa nucleoplasm (see Materials and Methods) after [^{35}S]methionine labeling in vivo; (B) hybrid selection with pHC12; (C) hybrid selection with the *NsiI-EcoRI* 3′ noncoding fragment which hybridizes only to the 1.9-kb poly(A)$^+$ RNA (see Fig. 1 and 3).

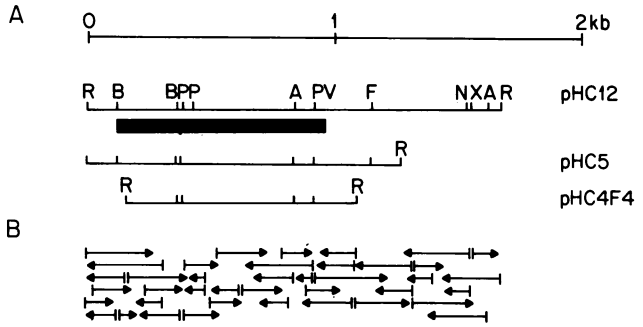

FIG. 3. Restriction map and sequencing strategy of HC4F4, HC5, and HC12. (A) Restriction maps of the C protein cDNAs. The poly(A) tracks at the 3′ end of HC5 and HC12 directly preceding the *EcoRI* sites are 12 and 10 bp, respectively. Abbreviations: A, *AvaII*; B, *BalI*; F, *FnuDII*; N, *NsiI*; P, *PstI*; PV, *PvuII*; R, *EcoRI*; X, *XbaI*. (B) DNA sequencing strategy. Sequencing was performed by the dideoxy chain termination method on fragments subcloned into M13mp18 and M13mp19. The solid bar indicates the protein coding region.

this mRNA or, as seems more likely, by posttranslational modification or processing of one by the other or by post-translational production from another primary translation product. This posttranslational possibility is also consistent with the sequence data discussed below, which indicate the presence in the mRNA of a single open reading frame. Furthermore, both proteins are also produced in a wheat germ translational system, indicating that the mechanism which generates the two polypeptides from the one (1.9-kb) mRNA is highly conserved (data not shown).

Nucleotide sequence of cDNA and deduced amino acid sequence of C proteins. The strategy used to determine the nucleotide sequence of HC12 is outlined in Fig. 3. The pHC12 insert, 1,686 nucleotides in size, contained a single uninterrupted open reading frame of 290 amino acids (Fig. 4). The putative initiation codon occurs within a strong consensus translational initiation sequence, AXXATGG (20), and the cDNA contains 114-bp 5′ and 687-bp 3′ untranslated sequences. The 3′ untranslated region contains two polyadenylation signals, AATAAA, about 400 bp apart. Protein sequencing of several peptides distributed throughout the protein has confirmed the deduced amino acid sequence (B. Merrill, K. Williams, and K. Schafer, personal communication).

The predicted amino acid sequence of the C proteins indicates that the primary translational product is 31,931 daltons, whereas the size estimated by mobility in SDS-PAGE varies for C1 between 37,000 and 42,000 and for C2 between 40,000 and 44,000 daltons (2, 5, 46). This discrepancy in molecular mass is probably due to the asymmetric charge distribution of the C proteins, particularly the very acidic carboxy domain, which may bind SDS poorly. The yeast transcriptional activator GCN4 also possesses an acidic stretch of amino acids followed by a basic carboxy terminus, and although it also migrates as a ~45,000-dalton protein in SDS-polyacrylamide gels, its actual size is 31,000 daltons (18). The amino acid composition indicates that the C proteins have a strong overall negative charge, consistent with the acidic isoelectric point observed by isoelectric focusing (12, 46). The C proteins have a unique and unusual distribution of charged and aromatic amino acids organized into two distinct domains. The carboxy part of the molecule is extremely acidic, containing long stretches of glutamic and aspartic acid residues but no aromatic amino acids or

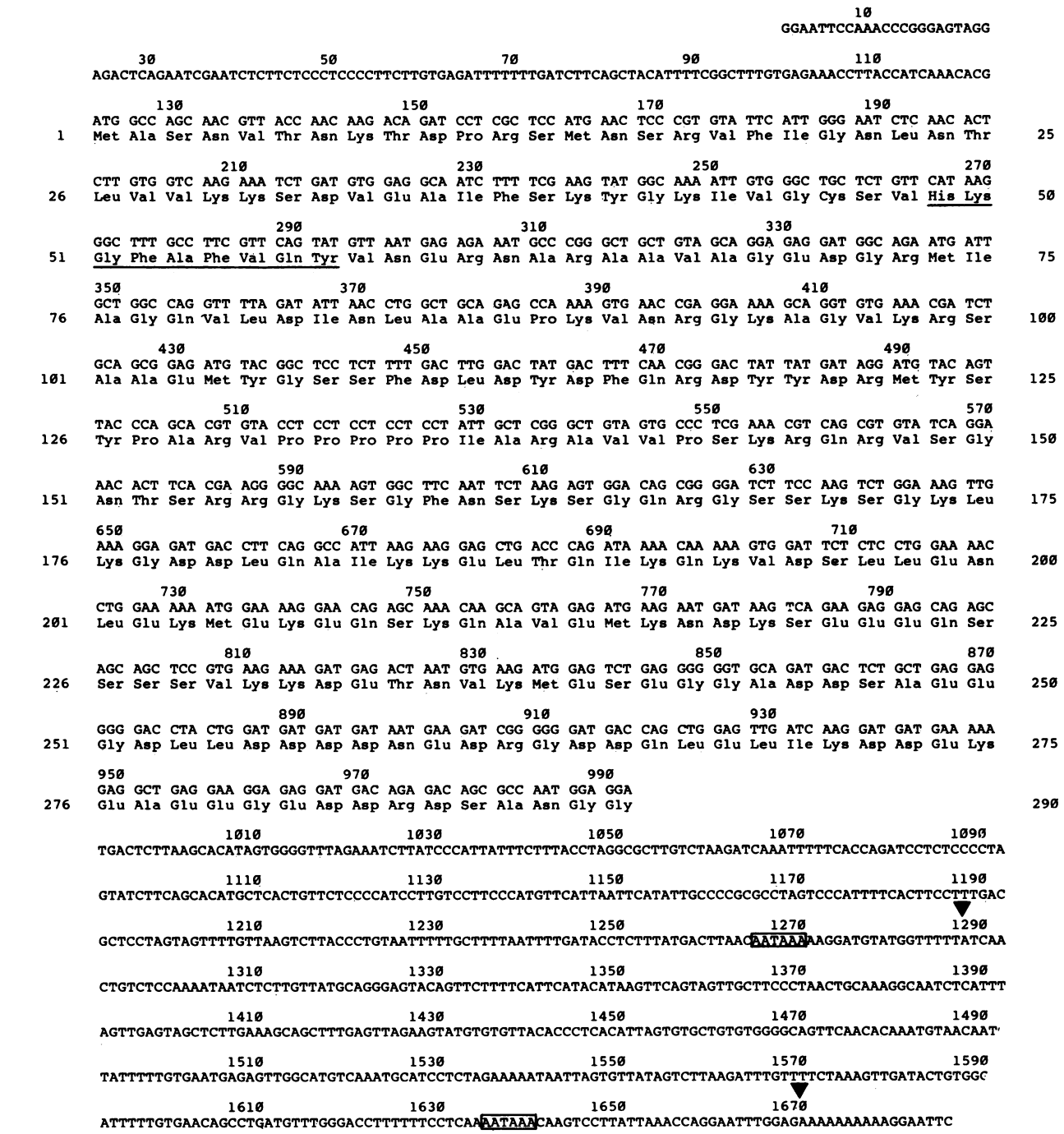

FIG. 4. Nucleotide sequence of the HC12 and HC5 cDNAs and the deduced amino acid sequence of the human C proteins. The RNP consensus sequence is underlined. The position of the poly(A) tail addition in HC5 and HC12 is indicated by arrowheads, and the polyadenylation signals are boxed.

prolines. The protein contains an RNP consensus sequence (1) in the amino-terminal domain beginning at about residue 49 (see below). The amino acid sequence is also rich in serine residues and contains a cluster of four serines at positions 225 to 228. This cluster is an expected phosphorylation site for a casein kinase type II (16), which has been shown to phosphorylate both C1 and C2 (17). Stretches of serines have also been found in the hnRNP protein A1 (21) and in the single-stranded nucleic-acid-binding protein of bacteriophage T4, the gene 32 protein (46a). We noted an (imperfect) inverted repeat sequence centered around Tyr104/Tyr105 amino terminal to a proline-rich domain within a cluster of aromatic and charged amino acids. One of the most interesting features of the protein structure is the presence of a putative NTP-binding fold. Comparison of the primary sequences of a number of ATPases has led to the suggestion that the consensus sequences Gly-$(X)_4$-Gly-Lys-(Thr/Gly)-$(X)_6$-(Ile/Val), referred to as the A site, and Arg-

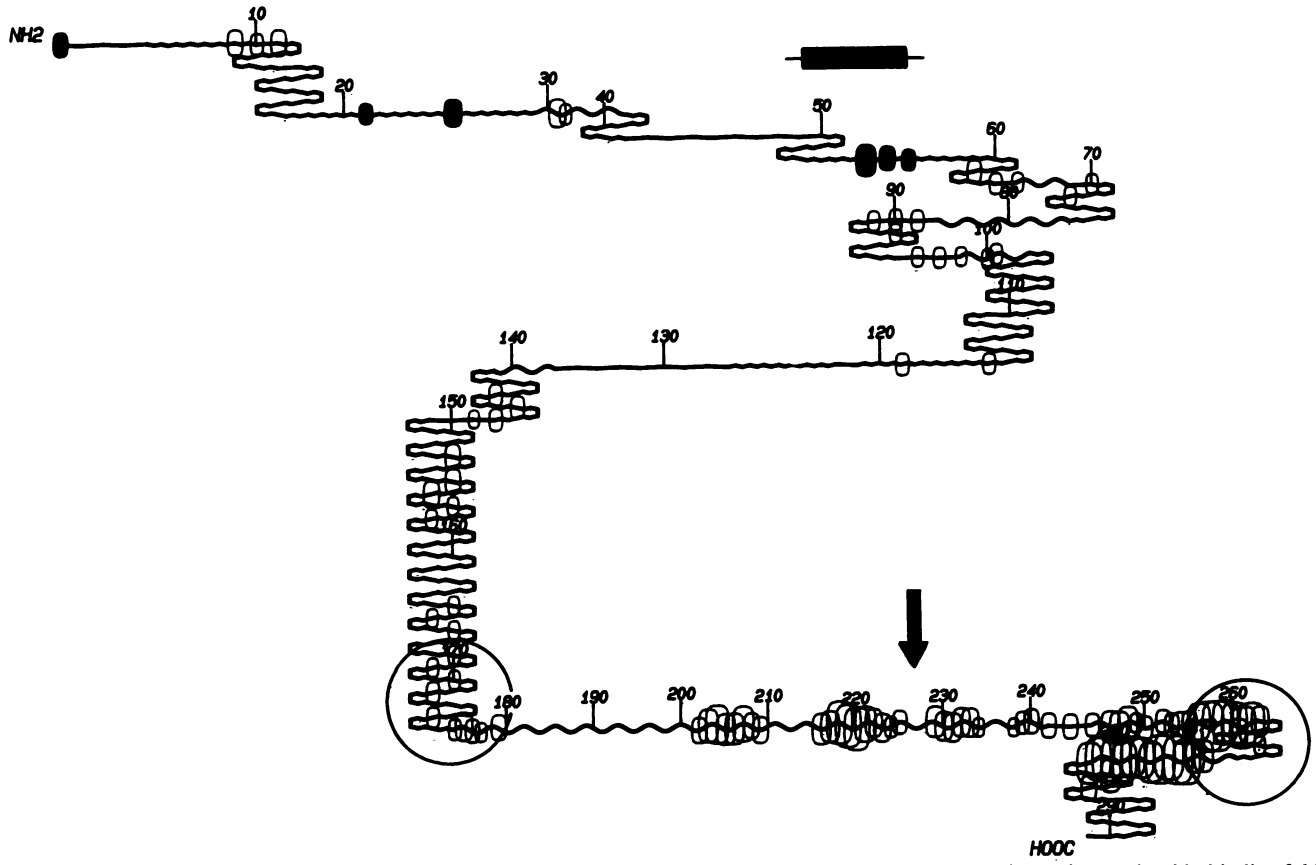

FIG. 5. Predicted secondary structure of the C protein, and position of the RNP consensus sequence, and putative nucleotide-binding fold. Secondary structures are predictions by the algorithm of Chou and Fasman (7). The positions of the RNP consensus sequence (▬), the ATPase consensus nucleotide-binding sequences (○), and casein kinase type II phosphorylation site (→) are marked. Hydrophobic moment (●) and charged residues (0) are shown with size proportional to magnitude. The predicted propensity for an α-helix formation (∿∿∿) and a β-pleated sheet (----) are indicated.

$(X)_{2-3}$-Gly-$(X)_3$-(hydrophobic)$_{4-6}$-Asp, the B site, constitute a nucleotide-binding fold (45). A Chou-Fasman prediction of the secondary structure (7) of the C proteins (Fig. 5) indicates that a similar pair of sequences occurs at either end of a sequence with the potential to form a long α-helical region between residues 168 and 183 (the A site) and 259 and 272 (the B site), whereas the putative phosphorylation site is in the middle of this segment (residues 225 to 228); another possible B site also exists near the amino terminus (residues 17 to 32). Altogether, the carboxy-terminal domain of the protein is very hydrophilic and is likely to be exposed on the surface of the protein. This is consistent with its accessibility to the protein kinase and perhaps to interaction with other proteins.

Probable production of the 1.9- and 1.4-kb C protein mRNAs from a single gene. As described above, the 1.9-kb RNA alone can translate in vitro into both C1 and C2. To understand the structure of the 1.4-kb mRNA, another clone (pHC5), which by restriction mapping overlapped with pHC12 at the 5′ end but not at the 3′ end, was sequenced. This clone has the same sequence as pHC12 (Fig. 3A), including the same 5′ end, except that the upstream polyadenylation signal was used, as indicated by the finding of a poly(A) segment 17 nucleotides 3′ to this signal (indicated by an arrowhead in Fig. 4). This information, together with the hybrid selection and in vitro translation data from fragments of pHC12 reported above, suggests that C1 and C2 are produced from the same mRNA and that the 1.4- and 1.9-kb poly(A)$^+$ RNAs are represented by pHC5 and pHC12, respectively. Production of multiple mRNAs from the same gene by differential use of polyadenylation sites has been observed for several other genes (e.g., 29, 44).

Conservation of RNP consensus sequence in hnRNA, mRNA and pre-rRNA-binding proteins. One of the most striking parts of the C protein sequence is Lys-Gly-Phe-Gly-Phe-Val-Thr-Tyr (Tyr57), since it is highly homologous to the RNP consensus sequence which we have recently described in the yeast mRNA poly(A)-binding protein (PABP) and the mammalian A1 hnRNP protein (1). We have now found a similar RNP consensus sequence also in the mammalian nucleolar 110-kDa pre-rRNA-binding protein C23 (24) and in the hnRNP protein UP2 (Fig. 6). The sequence of a partial cDNA "C" protein clone has recently been described by Lahiri and Thomas (22). This sequence is different from the one described here, which represents the C proteins as originally characterized by Beyer et al. (2) and as further defined with the monoclonal antibodies 4F4 and 2B12 (5, 12). Recent evidence indicates that the cDNA clone of Lahiri and Thomas codes for the UP2 protein (kindly provided by B. Merrill and K. Williams), which has similar mobility to the C proteins by SDS-PAGE but migrates differently in two-dimensional gel electrophoresis (M. Matunis and G. Dreyfuss, unpublished results). An RNP consensus se-

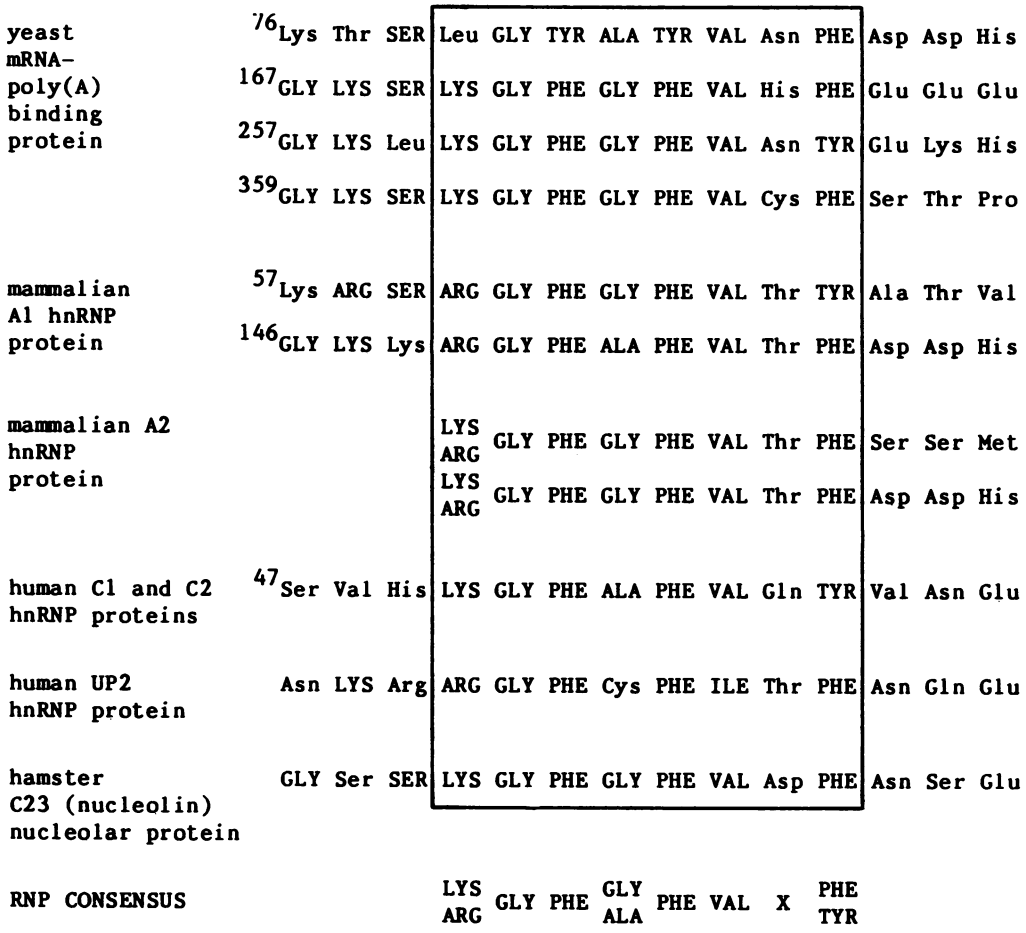

FIG. 6. RNP consensus sequence in RNA-binding proteins (capital letters in box). The numbers indicate the position of the amino acid in the protein. For C23, UP2, and A2, only partial sequences are available, and therefore the precise positions of the peptides in the proteins are unknown. Sources of information: for PABP, references 1 and 41; for A1, references 8 and 47; for A2, B. M. Merrill and K. R. Williams, personal communication; for C23, reference 24; and for UP2, reference 22. Because the sequence information for A2 is derived from tryptic peptides, the basic amino acid is either lysine or arginine.

quence if also found in tryptic fragments of the human and bovine A2 hnRNP protein (B. Merrill and K. Williams, personal communication), and it is likely to be present in the immunologically related (35) B1 and B2 proteins and in other RNP proteins. The first RNP consensus sequence in the PABP and the one in UP2 are somewhat more divergent than the others. Considerable sequence homology extends also beyond the RNP consensus sequence, over a domain of 80 to 100 amino acids (1, 41). The proteins also have an additional general feature in common: they all contain at least two distinct domains. The structures of the hnRNP proteins A1 and C1/C2, the mRNP PABP, and the nucleolar pre-rRNA-binding protein C23 (nucleolin) were compared on the basis of charge localization, Gly/Pro, Phe/Trp/Tyr, and the position of the RNP consensus sequence (Fig. 7). The amino-terminal region possesses either one (C proteins), two (A1), or four (PABP) RNP consensus sequences. Where there are multiple RNP consensus sequences, they appear to be at about the center of highly conserved repeats (~100 amino acids for the yeast PABP). Only partial sequence of the nucleolar C23 protein is available, and it contains at least one RNP consensus sequence amino-terminal to a glycine-rich domain similar to that found in A1. The carboxy termini of A1, PABP, and C23 are similar in that they possess a large number of both α-helix-breaking residues (glycine or proline) and aromatic amino acids, whereas the C proteins are extremely acidic, contain no aromatic residues, and (Fig. 5) may possess extensive α-helical character in this domain.

DISCUSSION

Although two mRNAs (1.9 and 1.4 kb) for the C proteins were detected by hybridization, the protein-coding capacity of both is the same. The difference between these two mRNAs is in the 3′ untranslated portion of the mRNA. Specific hybrid selection and translation of one of the mRNAs (the 1.9-kb mRNA; Fig. 2) demonstrated that indeed both proteins can be produced by translation of a single mRNA. The two mRNAs are therefore most likely produced from the same gene by differential use of two polyadenylation signals. It is likely that one C protein is produced from the other (e.g., C2 from C1) posttranslationally. It is less likely that two proteins are produced by differential use of the same mRNA, because only one open reading frame of sufficient length exists in the mRNA and because the putative initiation codon exists in a strong consensus sequence (20). The type of posttranslational modification or processing which might produce one C protein from the other is not

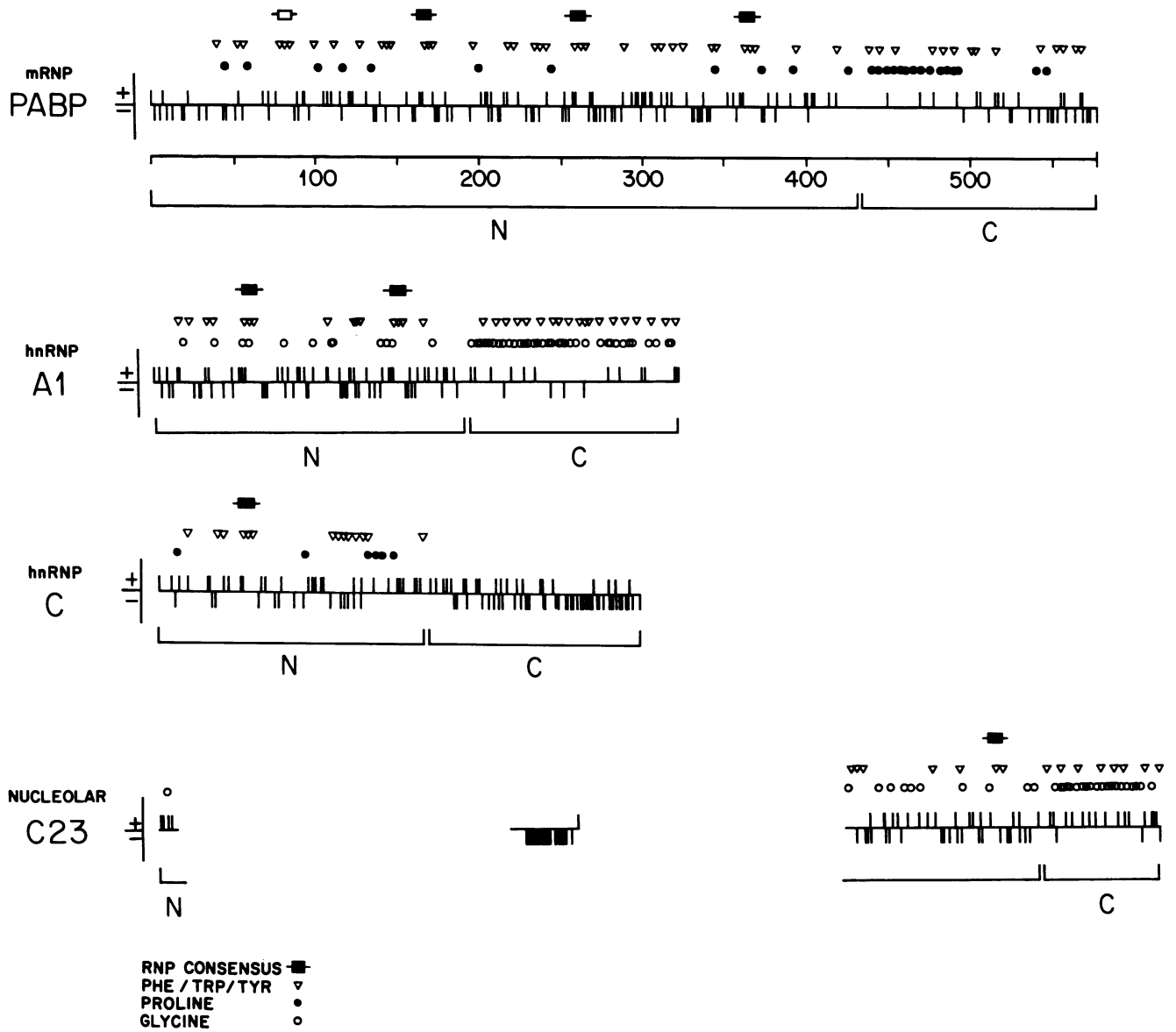

FIG. 7. Domain maps for mRNP PABP, hnRNP A1 and C1/C2 proteins, and nucleolar pre-rRNA-binding protein C23 (nucleolin). The sequence of C23 is based on partial amino acid and cDNA sequence information (24), and therefore the relative placement of peptides is arbitrary except for the carboxy terminus. Either proline or glycine positions are indicated, depending on which of these α-helix-disrupting residues is nonuniformly distributed throughout the protein.

known, but it must be highly conserved, since it is operational in both rabbit reticulocyte lysate and wheat germ extract. The amino acid compositions of C1 and C2 are very similar (see the corresponding C3 and C3X proteins described in reference 46), and Wilk et al. (46) have therefore raised the possibility that one C protein may be formed from the other by a specific proteolytic cleavage. It is intriguing that the ratio of C1 and C2 produced by translation in vitro of the 1.9-kb mRNA is similar to that found in vivo (Fig. 2). This process must therefore be extremely well regulated, since the ratio of C1 and C2 is always the same, C2 being one-third or one-fourth of C1 whether the proteins are produced in vivo or in vitro. It is possible, for example, that the primary translation product, possibly C1, forms oligomers, such as tetramers, and that one molecule per tetramer is modified to form C2. Phosphorylation is a known posttranslational modification of both C1 and C2 but does not appear to alter the C1/C2 ratio (17). Sachs et al. (41) have suggestd that the yeast nuclear PABP is derived by proteolytic cleavage of the cytoplasmic form of the protein, and the single-stranded DNA-binding protein UP1 has been shown to be a proteolysis product of the hnRNP A1 protein (21, 35, 40). The precedent for this type of posttranslational modification to produce a related but different RNA-binding protein is thus well established. Because many of the core proteins of the hnRNP appear by SDS-PAGE as doublets or higher multimers (e.g., the C proteins and the 68- and the 120-kDA proteins [4]), it is possible that the difference between C1 and C2 is a consequence of a distinct posttranslational modification or processing which affects the mobility of several of the polypeptides in SDS-PAGE and which is a hallmark of hnRNP proteins.

The sequence of the C proteins reveals several interesting features. First, the proteins contain the RNP consensus

sequence (1) which we have recently identified in the mRNA PABP and in the A1 hnRNP protein and which is also found in the major nucleolar pre-rRNA-binding protein C23 [24], in UP2 (22), and in the A2 hnRNP protein (B. Merrill and K. Williams, personal communication). It is interesting that C23 may be important in pre-rRNA processing and probably also interacts with itself and with other proteins to function as a nucleolar-organizing protein (34). The general and most highly conserved part of the consensus sequence is therefore Lys/Arg-Gly-Phe(Tyr)-Gly/Ala-Phe(Tyr)-Val(Ile)-X-Phe/Tyr, usually preceded by additional positively charged amino acids. It is therefore a cluster of interspersed aromatic amino acids preceeded by positively charged amino acids, and it forms a putative single-stranded RNA- and DNA-binding site within a relatively less conserved ~100-amino-acid domain. As discussed previously (1), it is likely that the aromatic amino acids interact with the nucleotide bases by hydrophobic stacking in which the Val is also important and that the positively charged residues interact electrostatically with the phosphate sugar backbone. It is likely that the sequence flanking the RNP consensus sequence in the RNA-binding domain, as well as other domains in the protein such as the carboxy section, determines the RNA-binding affinity and specificity. The RNP consensus sequence is not a universal feature of RNA-binding proteins; it is not found in ribosomal proteins and in the numerous known viral RNA-binding nucleocapsid proteins. Second, the C proteins have two distinct domains (Fig. 5): a very acidic carboxy terminus (residues 150 to 290) and an amino-terminal domain containing aromatic residues and the RNP consensus sequence. The negatively charged carboxy terminus of the C proteins possesses no aromatic residues or prolines. The other hnRNP protein for which sequence information is available, A1, as well as the mRNP yeast PABP, also possesses two discrete domains (1, 41). The amino-terminal segment contains either two (A1 and A2) or four (PABP) RNP consensus sequences within repeated domains. The carboxy half of these proteins possesses a large proportion of aromatic amino acids and is glycine rich in A1 and A2 and proline rich in the PABP. Therefore, since these proteins have an amino-terminal half with common sequence features, it is likely that this domain represents an RNA-binding site(s), because this is the one common functional activity displayed by all of these proteins. Furthermore, UP1, which is the 24-kDa amino-terminal portion of A1 without the glycine-rich carboxy domain (21, 40) (N of A1, Fig. 7), has single-stranded DNA- and RNA-binding activity (21, 40) and is composed of two repeating segments, each containing about 100 amino acids and one RNP consensus sequence (1). The RNP consensus sequence thus appears to be the most highly conserved part of an ancestral RNA-binding domain and is most likely part of the RNA-binding site. It is reasonable to anticipate that the carboxy domain of these RNA-binding proteins is important in their interaction with other proteins and in modulating the interaction of the RNA-binding domain with the polynucleotide. For the C proteins, the already extremely acidic carboxy domain contains a site for phosphorylation by a casein kinase type II. Phosphorylation increases the negative charge of this domain further and may have an important regulatory role. Third, the C protein sequence contains a putative NTP-binding fold and therefore may have the capacity to bind and possibly also hydrolyze ATP. This raises the interesting possibility that the C proteins may have the capacity to change conformation and perhaps to do work such as moving along RNA or having polynucleotide helicase activity. It is intriguing that the casein kinase II phosphorylation site on the C proteins appears to be in the middle of the sequence that must loop out for the putative NTP-binding fold to form.

ACKNOLWEDGMENTS

We thank Neal Farber of Biogen Inc. for providing the U937 λgt10 cDNA library, Robert Lamb for a gift of a wheat germ translational system, David Miller for computer and graphics plotter help, and Klaus Schafer, Barbara Merrill, and Kenneth Williams for generously communicating results of peptide sequencing.

This work was supported by Public Health Service grants GM-31888 and GM-37125 to G.D. from the National Institutes of Health.

ADDENDUM IN PROOF

A recent paper by Theissen et al. (EMBO J. **5**:3209–3217, 1986) reported the sequence of the 70-kDa UP1 snRNP protein. This snRNA-binding protein also contains an RNP consensus sequence, ^{316}LYS Pro ARG GLY TYR ALA PHE ILE Glu TYR.

LITERATURE CITED

1. **Adam, S. A., T. Nakagawa, M. S. Swanson, and G. Dreyfuss.** 1986. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. Mol. Cell. Biol. **6**:2932–2943.
2. **Beyer, A. L., M. E. Christensen, B. W. Walker, and W. M. LeStourgen.** 1977. Identification and characterization of the packaging proteins of core 40S hnRNP particles. Cell **11**:127–138.
3. **Billings, P. B., and T. E. Martin.** 1978. Proteins of nuclear ribonucleoprotein subcomplexes. Methods Cell Biol. **17**:349–376.
4. **Choi, Y. D., and G. Dreyfuss.** 1984. Isolation of the heterogeneous nuclear RNA ribonucleoprotein complex (hnRNP): a unique supramolecular assembly. Proc. Natl. Acad. Sci. USA **81**:7471–7475.
5. **Choi, Y. D., and G. Dreyfuss.** 1984. Monoclonal antibody characterization of the C proteins of heterogeneous nuclear ribonucleoprotein complexes in vertebrate cells. J. Cell Biol. **99**:1997–2004.
6. **Choi, Y. D., P. J. Grabowski, P. A. Sharp, and G. Dreyfuss.** 1986. Heterogeneous nuclear ribonucleoproteins: role in RNA splicing. Science **231**:1534–1539.
7. **Chou, P. Y., and G. D. Fasman.** 1978. Prediction of the secondary structure of proteins from their amino acid sequences. Adv. Enzymol. Rel. Areas Mol. Biol. **47**:45–148.
8. **Cobianchi, F., D. N. Sen Gupta, B. Z. Zmuda, and S. H. Wilson.** 1986. Structure of rodent helix-destabilizing protein revealed by cDNA cloning. J. Biol. Chem. **261**:3536–3543.
9. **Denhardt, D.** 1966. A membrane-filter technique for the detection of complementary DNA. Biochem. Biophys. Res. Commun. **23**:641–646.
10. **Dreyfuss, G.** 1986. Structure and function of nuclear and cytoplasmic ribonucleoprotein particles. Annu. Rev. Cell Biol. **2**: 457–496.
11. **Dreyfuss, G., S. A. Adam, and Y. D. Choi.** 1984. Physical change in cytoplasmic messenger ribonucleoproteins in cells treated with inhibitors of mRNA transcription. Mol. Cell. Biol. **4**:415–423.
12. **Dreyfuss, G., Y. D. Choi, and S. A. Adam.** 1984. Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. Mol. Cell. Biol. **4**:1104–1114.
13. **Economides, I. V., and T. Pederson.** 1983. Structure of ribonucleoproteins: heterogeneous nuclear RNA is complexed with a major sextet of proteins *in vivo*. Proc. Natl. Acad. Sci. USA **80**:1599–1602.
14. **Foe, V. E., L. E. Wilkinson, and C. D. Laird.** 1976. Comparative organization of active transcription units in *Oncopeltus fasciatus*. Cell **9**:131–146.

15. **Gall, J. G., and H. G. Callan.** 1962. ^{3}H-uridine incorporation in lampbrush chromosomes. Proc. Natl. Acad. Sci. USA **48:** 562–570.
16. **Hathaway, G. M., and J. A. Traugh.** 1982. Casein kinases—multipotential protein kinases. Curr. Top. Cell. Regul. **21:**101–127.
17. **Holcomb, E. R., and D. L. Friedman.** 1984. Phosphorylation of the C-proteins of HeLa cell hnRNP particles. J. Biol. Chem. **259:**31–40.
18. **Hope, I. A., and K. Struhl.** 1986. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. Cell **46:**885–894.
19. **Karn, J., G. Vidali, L. C. Boffa, and V. G. Allfrey.** 1977. Characterization of the non-histone nuclear proteins associated with rapidly labeled heterogeneous nuclear RNA. J. Biol. Chem. **252:**7307–7322.
20. **Kozak, M.** 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev. **47:** 1–45.
21. **Kumar, A., K. R. Williams, and W. Szer.** 1986. Purification and domain structure of core hnRNP proteins A1 and A2 and their relationship to single-strand DNA-binding proteins. J. Biol. Chem. **261:**11266–11273.
22. **Lahiri, D. K., and J. O. Thomas.** 1986. A cDNA clone of the hnRNP C proteins and its homology with the single-stranded DNA binding protein UP2. Nucleic Acids Res. **14:**4077–4094.
23. **Lamb, M. M., and B. Daneholt.** 1979. Characterization of active transcription units in Balbiani rings of *Chironemus tentans*. Cell **17:**835–848.
24. **Lapeyre, B., F. Amalric, S. H. Ghaffari, S. V. Venkatarama Rao, T. S. Dumbar, and M. O. J. Olson.** 1986. Protein and cDNA sequence of a glycine-rich, dimethylarginine-containing region located near the carboxy-terminal end of nucleolin (C23 and 100 kDa). J. Biol. Chem. **61:**9167–9173.
25. **Lehrach, H., D. Diamond, J. M. Wozney, and H. Boedtker.** 1977. RNA molecular weight determinations by gel electrophoresis under denaturing conditions, a critical re-examination. Biochemistry **16:**4743–4751.
26. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular Cloning: a laboratory manual (Cold Spring Harbor Laboratories, Cold Spring Harbor, N.Y.).
27. **Martin, T. E., P. Billings, A. Levey, S. Ozarsian, and I. Quinlan.** 1974. Some properties of RNA: protein complexes from the nucleus of eukaryotic cells. Cold Spring Harbor Symp. Quant. Biol. **38:**921–932.
28. **Martin, T. E., P. B. Billings, J. M. Pullman, B. J. Stevens, and A. J. Kinniburgh.** 1978. Substructures of nuclear ribonucleoprotein complexes. Cold Spring Harbor Symp. Quant. Biol. **42:**899–909.
29. **Masters, J. N., J. K. Yang, A. Cellini, and G. Attardi.** 1983. A human dihydrofolate reductase pseudogene and its relationship to the multiple forms of specific messenger RNA. J. Mol. Biol. **167:**23–36.
30. **McKnight, S. L., and O. L. Miller.** 1976. Ultrastructural patterns of RNA synthesis during early embryogenesis of *Drosophila melanogaster*. Cell **8:**305–319.
31. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101:**20–78.
32. **Miller, O. L., and A. H. Bakken.** 1972. Morphological studies of transcription. Karolinska Symp. Res. Meth. Reprod. Endocrinol. **5:**155–167.
33. **Nakagawa, T. Y., M. S. Swanson, B. J. Wold, and G. Dreyfuss.** 1986. Molecular cloning of cDNA for the nuclear ribonucleoprotein particle C proteins: a conserved gene family. Proc. Natl. Acad. Sci. USA **83:**2007–2011.
34. **Orrick, L. R., O. J. Olson, and H. Busch.** 1973. Comparison of nucleolar proteins of normal rat liver and Novikoff hepatoma ascites cells by two-dimensional polyacrylamide gel electrophoresis. Proc. Natl. Acad. Sci. USA **70:**1316–1320.
35. **Pandolfo, M., O. Valentini, G. Biamonti, C. Morandi, and S. Riva.** 1985. Single-stranded DNA binding proteins derive from hnRNP proteins by proteolysis in mammalian cells. Nucleic Acids Res. **13:**6577–6590.
36. **Pederson, T.** 1974. Proteins associated with heterogeneous nuclear RNA in eukaryotic cells. J. Mol. Biol. **83:**163–183.
37. **Pelham, H. R. B., and R. J. Jackson.** 1976. An efficient mRNA-dependent translation system from reticulocyte lysates. Eur. J. Biochem. **67:**247–256.
38. **Ricciardi, R. P., J. S. Miller, and B. E. Roberts.** 1979. Purification and mapping of specific mRNAs by hybridization selection and cell-free translation. Proc. Natl. Acad. Sci. USA **76:**4927–4931.
39. **Rigby, P. W. J., M. Dieckmann, C. Rhodes, and P. Berg.** 1977. Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113:**237–251.
40. **Riva, S., C. Morandi, P. Tsoulfus, P. Pandolfo, M. Biamonti, B. Merrill, K. R. Williams, G. Multhany, K. Bayreuther, H. Werr, B. Heinrich, and K. P. Schaffer.** 1986. Mammalian single-stranded DNA binding protein UP1 is derived from the hnRNP protein A1. EMBO J. **5:**2267–2274.
41. **Sachs, A. B., W. M. Bond, and R. D. Kornberg.** 1986. A single gene from yeast for both nuclear and cytoplasmic polyadenylate-binding proteins: domain structure and expression. Cell **45:** 827–835.
42. **Samarina, O. P., and A. A. Krichevskaya.** 1981. Nuclear 30S RNP particles, p. 1–48. In H. Busch (ed.), The Cell Nucleus, vol. 9. Academic Press, Inc., New York.
43. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
44. **Scarpulla, R. C.** 1984. Processed pseudogenes for rat cytochrome *c* are preferentially derived from one of three alternate mRNAs. Mol. Cell. Biol. **4:**2279–2288.
45. **Walker, J. E., M. Saraste, M. J. Runswick, and N. J. Gay.** 1982. Distantly related sequences in the α and β subunits of ATP synthetase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide fold. EMBO J. **1:**945–951.
46. **Wilk, H., H. Werr, D. Friedrich, H. H. Klitz, and K. P. Schafer.** 1985. The core proteins of the 35S hnRNP complex. Eur. J. Biochem. **146:**71–81.

46a. **Williams, K. R., M. B. LoPresti, M. Setoguchi, and W. H. Konigsberg.** 1980. Amino acid sequence of the T4 DNA helix-destabilizing protein. Proc. Natl. Acad. Sci. USA **77:**4614–4617.

47. **Williams, K. R., K. L. Stone, M. B. LoPresti, B. M. Merrill, and S. R. Planck.** 1985. Amino acid sequence of the UP1 calf thymus helix-destabilizing protein and its homology to an analogous protein from mouse myeloma. Proc. Natl. Acad. Sci. USA **82:** 5666-5670.