The EMBO Journal vol.11 no.7 pp.2655-2664, 1992

Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box

Megerditch Kiledjian and Gideon Dreyfuss

Howard Hughes Medical Institute, and Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6148, USA

Communicated by W.Keller

Heterogeneous nuclear ribonucleoproteins (hnRNPs) are thought to influence the structure of hnRNA and participate in the processing of hnRNA to mRNA. The hnRNP U protein is an abundant nucleoplasmic phosphoprotein that is the largest of the major hnRNP proteins (120 kDa by SDS-PAGE). HnRNP U binds pre-mRNA *in vivo* and binds both RNA and ssDNA *in vitro*. Here we describe the cloning and sequencing of a cDNA encoding the hnRNP U protein, the determination of its amino acid sequence and the delineation of a region in this protein that confers RNA binding. The predicted amino acid sequence of hnRNP U contains 806 amino acids (88 939 Daltons), and shows no extensive homology to any known proteins. The N-terminus is rich in acidic residues and the C-terminus is glycine-rich. In addition, a glutamine-rich stretch, a putative NTP binding site and a putative nuclear localization signal are present. It could not be defined from the sequence what segment of the protein confers its RNA binding activity. We identified an RNA binding activity within the C-terminal glycine-rich 112 amino acids. This region, designated U protein glycine-rich RNA binding region (U-gly), can by itself bind RNA. Furthermore, fusion of U-gly to a heterologous bacterial protein (maltose binding protein) converts this fusion protein into an RNA binding protein. A 26 amino acid peptide within U-gly is necessary for the RNA binding activity of the U protein. Interestingly, this peptide contains a cluster of RGG repeats with characteristic spacing and this motif is found also in several other RNA binding proteins. We have termed this region the RGG box and propose that it is an RNA binding motif and a predictor of RNA binding activity.

Key words: nuclear proteins/RGG box/RNA binding motif/RNP

Introduction

Heterogeneous nuclear ribonucleoproteins (hnRNPs) are among the most abundant proteins in the eukaryotic cell nucleus. They associate with nascent RNA polymerase II transcripts to form hnRNP complexes and are thought to influence the structure of hnRNA and participate in pre-mRNA processing (reviewed in Dreyfuss, 1986; Dreyfuss *et al.*, 1988). Monoclonal antibodies to several hnRNP proteins have been used to immunopurify hnRNP complexes from vertebrate cells and have demonstrated that they contain at least 20 abundant proteins ranging in size from 34 kDa (A1) to 120 kDa (U) (Choi and Dreyfuss, 1984a; Piñol-Roma *et al.*, 1988). The 120 kDa hnRNP U protein is an abundant nucleoplasmic phosphoprotein (Dreyfuss *et al.*, 1984a) that can be crosslinked to pre-mRNA in intact cells by UV light (Dreyfuss *et al.*, 1984b). It is co-immunopurified with antibodies to other hnRNP proteins indicating that it is part of the same supramolecular complexes that contain the other hnRNP proteins (Choi and Dreyfuss, 1984b; Piñol-Roma *et al.*, 1988).

To better understand the function of the hnRNP U protein and its influence on the fate of pre-mRNA, we identified a cDNA clone encoding it and characterized the RNA binding properties of the protein. The sequence predicts an 806 amino acid protein. The N-terminal portion has an acidic stretch of amino acids followed by a glutamine-rich region. The C-terminus is rich in glycine, asparagine and arginine. However, the sequence of U does not contain a canonical consensus sequence RNA binding domain (RBD) as has been found in many of the other hnRNP proteins (see Dreyfuss *et al.*, 1988; Bandziulis *et al.*, 1989; Kenan *et al.*, 1991 for reviews). We therefore determined the RNA binding region of the U protein by deletional analysis and found that the C-terminal 112 amino acid glycine-rich segment is necessary and sufficient for RNA binding. Fusion of this U protein glycine-rich RNA binding region (U-gly) to the bacterial maltose binding protein converts this non-nucleic acid binding protein into an RNA binding protein. The RNA binding activity within U-gly was further localized to a 26 amino acid region which contains a cluster of RGG repeats. This region, termed the 'RGG box', is necessary for RNA binding of the U protein. Interestingly, an RGG box is found in several other RNA binding proteins including the nucleolar proteins SSB-1 (Jong *et al.*, 1987), nucleolin (Lapeyre *et al.*, 1987; Bourbon *et al.*, 1988; Srivastava *et al.*, 1989; Caizergues-Ferrer *et al.*, 1989; Maridor *et al.*, 1990), and fibrillarin (Henriquez *et al.*, 1990; Aris and Blobel, 1991) and the hnRNP A1 protein (Cobianchi *et al.*, 1986; Buvoli *et al.*, 1988).

Results

Isolation of cDNA clones encoding the hnRNP U protein

cDNA clones encoding the hnRNP U protein were identified by immunoscreening of a HeLa λZap II library with the monoclonal antibody, 3G6, which specifically recognizes the U protein (Dreyfuss *et al.*, 1984b). One immunoreactive clone of ~1 kb was identified and used to isolate three additional clones by hybridization screening. The largest clone, U21.1, containing a 3.2 kb insert, was used to generate [^{35}S]methionine-labeled protein by *in vitro* transcription and translation. The *in vitro* produced protein co-migrated with authentic 3G6-immunopurified U protein from HeLa cells by SDS-PAGE (Figure 1A, lanes 1 and 2). Several additional criteria were used to verify that the

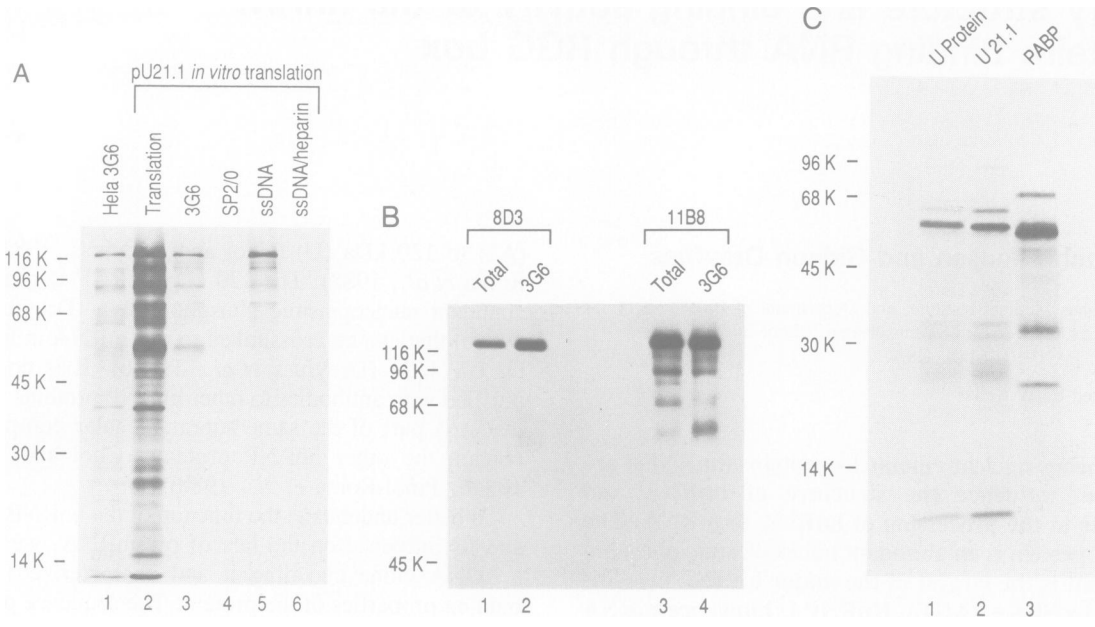

Fig. 1. U21.1 clone encodes the hnRNP U protein. (**A**) Lane 1 is anti-U protein monoclonal antibody, 3G6, immunopurified U protein from [^{35}S]methionine-labeled HeLa cells. The U21.1 cDNA was transcribed and translated *in vitro* (lane 2) and immunoprecipitated with 3G6 (lane 3) or the non-immune antibody SP2/0 (lane 4). The U21.1 gene product was bound to ssDNA−agarose beads and washed in the same buffer without heparin (lane 5) or with 2 mg/ml heparin (lane 6). (**B**) U protein in total HeLa cell extract (lanes 1 and 3) or 3G6 immunopurified U protein from HeLa cell extract (lanes 2 and 4) was immunoblotted with monoclonal antibodies 8D3 and 11B8 generated against the U21.1 clone. (**C**) Peptide maps of 3G6 immunopurified U protein from HeLa cells (lane 1) and U21.1 translation product (lane 2) were generated by cleavage at tryptophan residues with the chemical agent BNPS-Skatole. Poly(A) binding protein (PABP) of *S.cerevisiae* was treated with the same reagent as a control (lane 3).

U21.1 clone was a cDNA for the hnRNP U protein. *In vitro* translated U21.1 was immunoprecipitated with 3G6, but not with the non-specific immunoglobulin from the parental myeloma cell line SP2/0 (Figure 1A, compare lanes 3 and 4). As demonstrated in lanes 5 and 6, the U21.1-encoded protein can bind ssDNA and the binding is sensitive to heparin. Both of these properties are consistent with properties of the U protein purified from HeLa cells (Piñol-Roma *et al.*, 1988). Furthermore, monoclonal antibodies raised against bacterially produced U21.1-encoded protein recognize the 120 kDa protein immunopurified with 3G6 (Figure 1B).

We devised a refined tryptophan cleavage method to generate a peptide map of the U21.1 protein product to verify its identity with the U protein further. [^{35}S]methionine-labeled U protein was immunopurified with 3G6 from either HeLa cells or from an *in vitro* translation product of the U21.1 clone. The proteins were then resolved by SDS−PAGE, blotted onto Immobilon-P membrane and cleaved *in situ* at tryptophan residues with BNPS-Skatole (Omenn *et al.*, 1970). As shown in Figure 1C the resulting peptides are identical for both proteins (lanes 1 and 2). The poly(A) binding protein of *Saccharomyces cerevisiae* (Adam *et al.*, 1986) treated with the same reagent was included as a control (lane 3). This mapping method as described here is simple and rapid, and could be useful for a wide range of peptide mapping applications. The facts that U21.1-encoded protein is recognized by 3G6, antibodies raised against it recognize 3G6-immunopurified HeLa U protein, and it has an identical peptide map to the U protein, strongly indicate that the U21.1 cDNA encodes the full length hnRNP U protein.

Primary structure of the U protein

The nucleotide sequence of U21.1 cDNA and predicted amino acid sequence of the U protein as well as a schematic representation of the protein are shown in Figure 2. The U21.1 cDNA is 3223 bp and it contains an open reading frame that encodes a predicted 806 amino acid protein with a calculated molecular mass of 88 939 Daltons and a pI of 5.5. The proposed initiation codon at nucleotide 41 contains considerable homology to the translation initiation sequence consensus (Kozak, 1983) and is preceded by two in-frame stop codons. The clone extends 791 nucleotides 3′ of the translation stop codon but it may not contain the entire 3′ untranslated region since a poly(A) tract is not present. Nucleotide and protein sequence searches of the Genbank and EMBL databases (Lipman and Pearson, 1985) did not reveal extensive homology to any known sequence. The N-terminal 160 amino acids of the protein are rich in acidic amino acids. Of the amino acids in this region 33% are aspartic and glutamic acid. Within the next 50 amino acids, 28% are glutamine residues followed by a putative nuclear localization signal (Figure 2). A GX_2GXGKT consensus sequence for a putative NTP binding site (Walker *et al.*, 1982) is found at amino acids 485−492. A secondary structure prediction (Chou and Fasman, 1978) of this region is consistent with the consensus sequence positioned within the turn portion of a Rossmann-fold (Rossmann *et al.*, 1974) between a β-strand and an α-helix. As is the case with several other hnRNP proteins, the C-terminal region is rich in glycine residues, having 26% glycine in the C-terminal 129 amino acids. The predicted protein appears to contain multiple potential casein and histone kinase phosphorylation sites as well as several N-linked glycosylation consensus

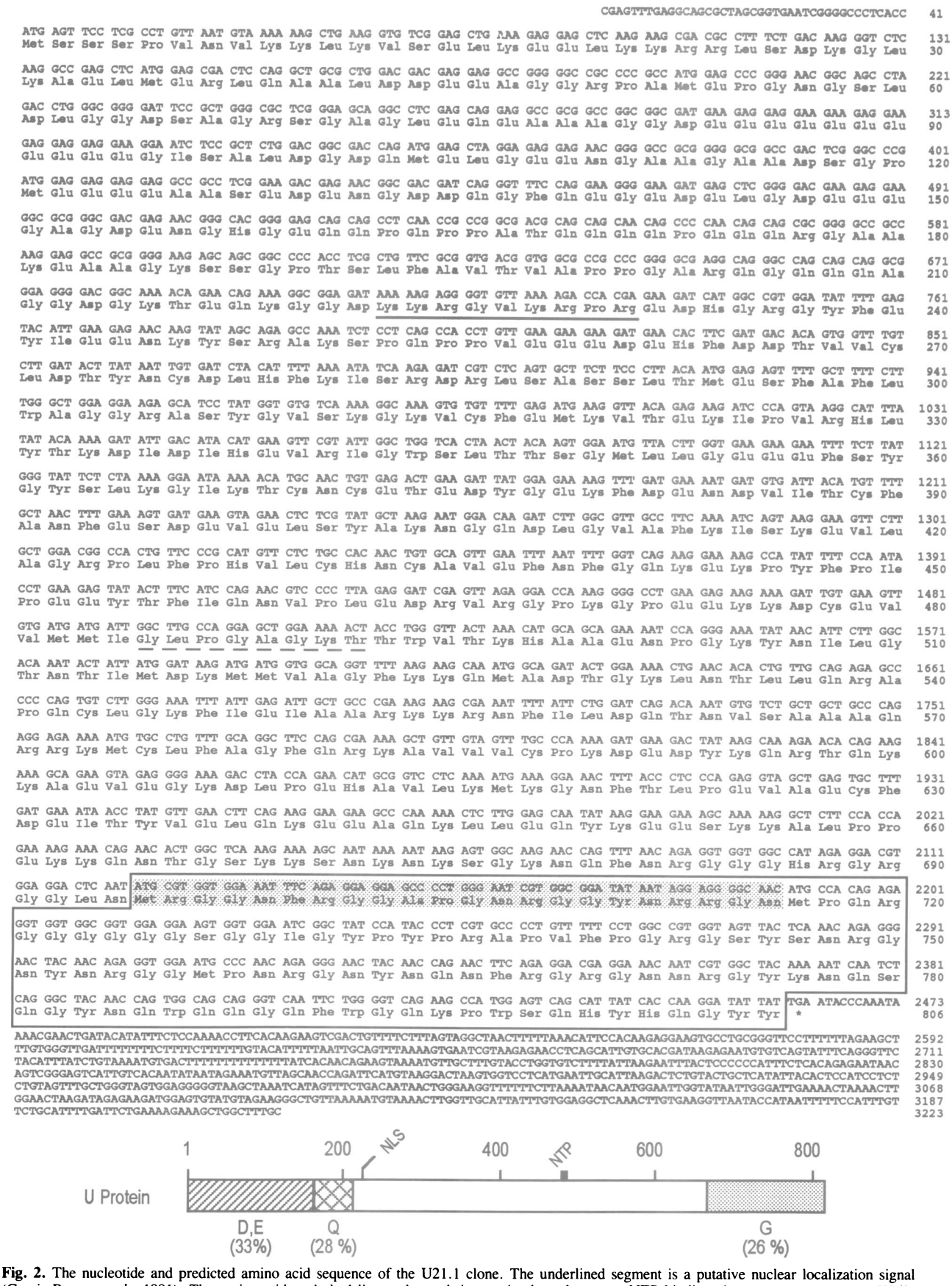

Fig. 2. The nucleotide and predicted amino acid sequence of the U21.1 clone. The underlined segment is a putative nuclear localization signal (Garcia-Bustos *et al.*, 1991). The region with a dashed line underneath it contains homology to a NTP binding site consensus sequence (Walker *et al.*, 1982). The boxed region is the glycine-rich RNA binding region (U-gly), and the shaded area is the RGG box. Below the sequence is a schematic representation of the U protein with the amino acid numbers above it. NLS and NTP denote the putative nuclear localization and NTP binding sites. The identity of the shaded areas along with the percentage of the indicated amino acids are as indicated. D and E are aspartic acid and glutamic acid respectively, Q is glutamine and G is glycine.

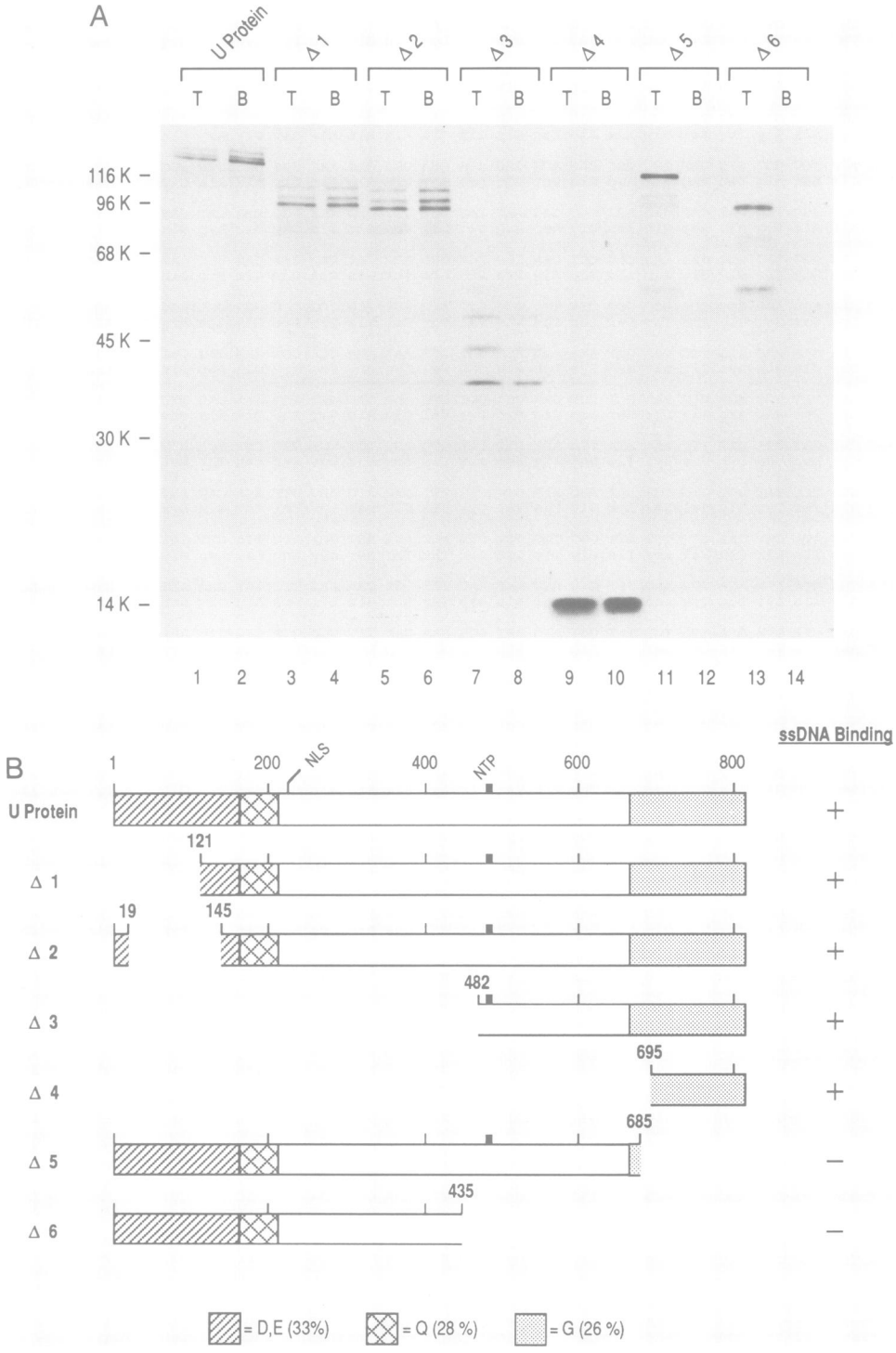

Fig. 3. The single-stranded nucleic acid binding domain is located at the C-terminal end of the U protein. **(A)** All DNAs were linearized with *Hin*dIII and used as template for SP6 RNA polymerase transcription and *in vitro* rabbit reticulocyte lysate translation. *In vitro* produced protein was bound to 30 μg ssDNA–agarose beads at 250 mM NaCl and analyzed by SDS–PAGE as described in Materials and methods. The 'T' lanes contain a total translation product equivalent to 30% of the indicated protein used in the binding reaction ('B' lanes). The mol. wt markers (in kDa) are indicated on the side of the gel. The identities of the protein-deletion constructs are illustrated in (B) and were derived from the following plasmids: U protein (pGem-U21.1); Δ1 (pGem-5′ΔApa); Δ2 (pGem-IΔSst); Δ3 (pGem-5′ΔBgl); Δ4 (pGem-5′ΔMsc); Δ5 (pGemΔ3′Msc); Δ6 (pGemΔ3′Acc). **(B)** The U protein and various deleted forms of the protein are schematically depicted with the amino acid numbers shown above them. The identities of the shaded areas are shown in the key at the bottom and described in the legend to Figure 2. The binding activity of the various deletion constructs from panel A are summarized as '+' or '–' for those that bind or do not bind respectively.

sequences. The mobility of U protein on two-dimensional gels (Piñol-Roma *et al.*, 1988) is consistent with the predicted acidic nature of the protein with many potential phosphorylation sites. A slower mobility by SDS–PAGE than expected from the actual molecular weight (for hnRNP U ~89 kDa and ~120 kDa respectively) has been observed for other acidic and phosphorylated proteins (Hope and Struhl, 1986; Swanson *et al.*, 1987).

Identification of the U protein single-stranded nucleic acid binding domain

HnRNP U is bound to pre-mRNA *in vivo* (Dreyfuss *et al.*, 1984b) and binds RNA and ssDNA *in vitro* (Piñol-Roma *et al.*, 1988). However, unlike many hnRNP proteins, the U protein does not contain a consensus sequence RNA binding domain (RBD). This prompted us to investigate what region of the U protein confers RNA binding. Deletions were therefore constructed to generate *in vitro* translated U protein fragments having various N-terminal, internal and C-terminal truncations. The intact protein and protein fragments were bound to ssDNA–agarose beads as described in Materials and methods and analyzed by SDS–PAGE. The translation product of the N-terminal truncations results in multiple bands (Figure 3A, lanes 1, 3, 5 and 7). As the U protein is heavily phosphorylated, the more slowly migrating polypeptides probably result from phosphorylation. As shown in Figure 3A (and summarized in 3B), all the deletions that retain the C-terminus are capable of binding ssDNA (lanes 1–10, compare translation 'T' lanes with bound 'B' lanes). However, removal of the C-terminal segment of U protein renders it unable to bind single-stranded nucleic acid (lanes 11–14). All detected binding was to the ssDNA as no binding was detected to the support matrix (data not shown). Therefore, the strongest (and possibly only) single-stranded nucleic acid binding activity of the U protein resides in the C-terminal 112 amino acids. Using less stringent binding conditions (lower salt concentrations) an additional weaker binding activity was found in the remainder of the protein (data not shown). For the purpose of this paper, we will limit our discussion only to the strongest binding activity of the U protein contained at the C-terminus to which we refer as U protein glycine-rich RNA binding region (U-gly). In Figure 4A the binding pattern of U protein to ribonucleotide homopolymers at salt concentrations ranging from 0.25 M to 1 M NaCl is shown. The binding pattern of U-gly (panel B) parallels that of the U protein. Both the U protein and U-gly have the highest salt-resistant binding to poly(G), intermediate binding to poly(A) and poly(U), and very weak binding to poly(C). The intact U protein and U-gly thus have similar RNA binding properties. Therefore, U-gly is the RNA binding segment of the U protein.

The C-terminus of the hnRNP A1 protein is also rich in glycine (A1-gly) and has been reported to have RNA binding activity at low salt concentrations (Cobianchi *et al.*, 1988; Nadler *et al.*, 1991). We therefore compared the binding efficiency of A1-gly with that of U-gly. The C-terminal 135 amino acids of A1 were produced by translation *in vitro* and binding experiments to ssDNA were carried out at various salt concentrations side by side with U-gly. Figure 5 shows the comparison of A1-gly binding with that of U-gly binding under identical conditions. U-gly bound at salt concentrations up to 500 mM NaCl, while the binding of A1-gly was completely abolished at NaCl concentrations above 100 mM. Thus while both glycine-rich regions have a similar, unusually high, glycine content, they appear to bind single-stranded nucleic acid with different characteristics; U-gly binding appears to be a much stronger salt-resistant single-stranded nucleic acid binding peptide than A1-gly.

To determine if U-gly is sufficient for RNA binding activity we tested its ability to confer this property onto an otherwise non-nucleic acid binding protein. A plasmid was constructed that encodes a fusion protein having the bacterial maltose binding protein with U-gly at the C-terminus. The translation product of the MalU3′gly protein had the predicted size by SDS–PAGE and the identity of the predicted fusion protein encoding construct was confirmed

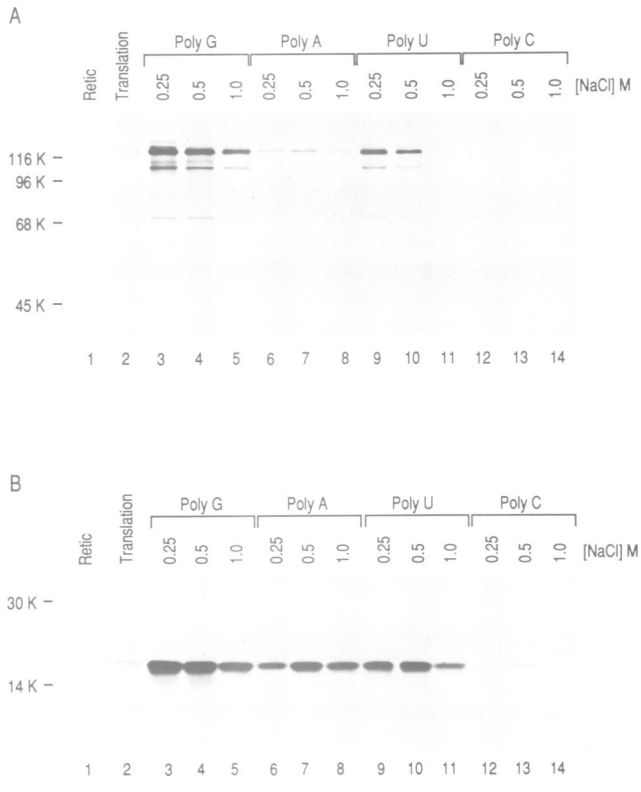

Fig. 4. U-gly contains the same RNA binding specificity as the U protein. (**A**) Full-length *in vitro* translated U21.1 protein was bound to 30 μg of the indicated ribonucleotide homopolymers at the indicated salt concentrations as described in the legend to Figure 3. Lane 1 represents unprogrammed reticulocyte lysate translation. Lane 2 is an equivalence of one-tenth the amount of U21.1 translation product used in the bound lanes. Molecular weight markers are indicated at the edge of the panel. (**B**) Binding of *in vitro* produced glycine-rich RNA binding region (U-gly) to ribonucleotide homopolymers as indicated above.

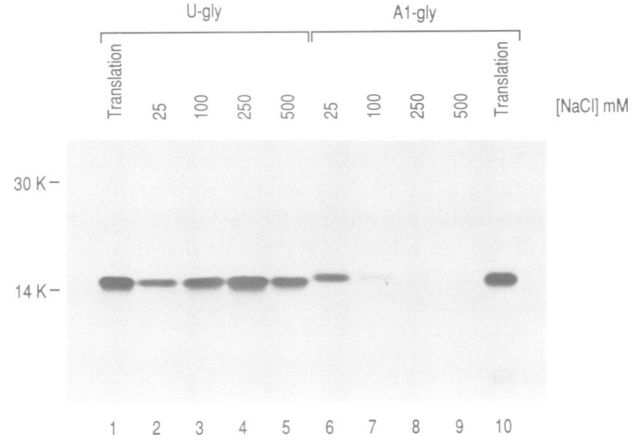

Fig. 5. U-gly binds ssDNA with higher salt resistance than A1-gly. Equal counts of *in vitro* translated U-gly or A1-gly (amino acids 186–320 of the hnRNP A1 protein) were bound to 30 μg ssDNA at the indicated NaCl concentrations, eluted from the beads and analyzed by SDS–PAGE. Lanes 1–5 are binding of U-gly and lanes 6–10 are binding of A1-gly . Lanes 1 and 10 are aliquots of total translation product equivalent to 40% of the total amount of protein used in the binding reactions for U-gly or A1-gly respectively.

by DNA sequencing. As shown in Figure 6, maltose binding protein alone does not bind ssDNA−agarose or ribonucleotide homopolymer−agarose beads (lanes 5−8). However, when fused to U-gly, it can bind both ssDNA and RNA (lanes 1−4). Similar results were observed using the bacterially produced fusion protein (data not shown). As would be expected for an hnRNA binding protein, the observed interaction with RNA was more efficient than that with ssDNA. Therefore, U-gly is an autonomous RNA binding peptide that contains an RNA binding domain.

Having established that U-gly is an RNA binding polypeptide we set out to identify the smallest single-stranded nucleic acid binding unit within it. Due to the difficulty in translating small peptides *in vitro*, the U-gly deletion cDNAs were constructed such that the synthesized polypeptides contained an additional 20 amino acids encoded by the pGem4 polylinker (see Materials and methods). Binding activity (percentage bound relative to total input protein) of U-gly with the C-terminal 10 amino acids substituted by the polylinker-encoded 20 amino acids (U-glyΔ1) was similar to the binding activity of U-gly alone (data not shown). Subsequent deletions were compared with U-glyΔ1. Removal of the 49 C-terminal amino acids of U-gly had a slight effect on the peptide's ability to bind single-stranded nucleic acid, while a truncation at the N-terminus abolished binding (Figure 7A). The fused 20 amino acid peptide from the pGem4 polylinker does not contain RNA binding activity since the N-terminal truncation also includes this peptide and does not bind. Thus, a 63 amino acid peptide between amino acids 695 and 757 retains the binding characteristics of U-gly.

We were unable efficiently to translate *in vitro* smaller C-terminal truncated U-gly peptides and, therefore, additional truncations were generated within the full length U protein. As shown in Figure 7B, a truncation of the U protein C-terminus up to amino acid 720 (Δ8) does not interfere with the binding of this protein to single-stranded nucleic acids while removal of sequences to amino acid 685 abolished binding (Δ5, also see Figure 3). A schematic representation of these data is shown in panel C. Comparison of the N- and C-terminal deleted proteins that still retain binding activity revealed a 26 amino acid sequence from 695 to 720 that is present in both. Furthermore, this peptide is clearly necessary for binding activity because its removal from U-gly abolishes binding to RNA (Figure 7A, compare lane 4 with 6). Therefore, it is possible that this 26 amino acid peptide constitutes the entire RNA binding domain of the protein.

Discussion

We describe here the cDNA cloning and sequencing, and characterization of the RNA binding activity, of the hnRNP U protein. HnRNP U, with an apparent molecular mass of 120 kDa by SDS−PAGE, is the largest of the abundant hnRNP proteins. The predicted protein sequence reveals acidic and glutamine-rich regions at the N-terminus, a putative NTP binding site, a candidate nuclear localization signal and a glycine-rich C-terminus. It is intriguing that nucleolin, the major 110 kDa nucleolar pre-rRNA binding protein, also has acidic and glycine-rich domains at the N- and C-termini, respectively, although, unlike U, it contains four RBDs (Lapeyre *et al.*, 1987). Although the U protein is an RNA binding protein, it does not contain significant

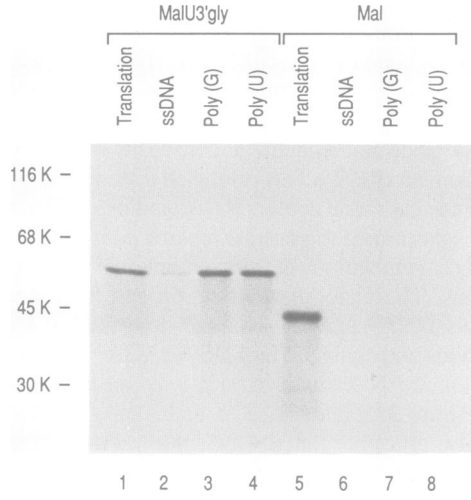

Fig. 6. U-gly can convert the maltose binding protein into an RNA binding protein. *In vitro* produced maltose binding protein fused to the C-terminal 122 amino acids of the U protein, (MalU3'gly, this includes 10 additional amino acids N-terminal to U-gly alone), or maltose binding protein alone (Mal) were bound to ssDNA (lanes 2 and 6), poly(G) (lanes 3 and 7) or poly(U) (lanes 4 and 8). Lanes 1 and 5 are the respective translation products equivalent to 30% used in the bound lanes. Binding reactions were with 30 μg nucleic acid on agarose beads.

amino acid sequence homology to any previously identified RNA binding motif. We have identified the C-terminal glycine-rich region of the protein as the most avid RNA binding region of the protein. This segment of the protein, termed U protein glycine-rich RNA binding region (U-gly), has the same RNA binding characteristics to several synthetic RNA and DNA substrates as the entire U protein, and its removal abolishes the binding activity of the protein at 250 mM NaCl.

Several hnRNP proteins, including the human hnRNP A1 and A2 proteins as well as numerous other hnRNP proteins from diverse organisms, also have a glycine-rich region (Cobianchi *et al.*, 1986; Burd *et al.*, 1989; Kay *et al.*, 1990; Matunis *et al.*, 1992a). In the case of the hnRNP A1 protein, the C-terminal region is comprised of 45% glycine and has been suggested to be involved in protein−protein and protein−nucleic acid interactions (Cobianchi *et al.*, 1988; Nadler *et al.*, 1991). A synthetic peptide containing 46 amino acids of the glycine-rich region of the hnRNP A1 protein binds RNA at NaCl concentrations of 10−25 mM (Nadler *et al.*, 1991) which is consistent with the binding of the entire carboxyl A1-gly region to ssDNA shown in this work (Figure 5). In comparison, U-gly, which has a 27% glycine content, is a much more efficient salt-resistant single-stranded nucleic acid binding protein capable of binding ssDNA at 500 mM NaCl. Therefore, U-gly appears to bind single-stranded nucleic acid *in vitro* much more tightly than the previously identified A1-gly region.

Further delineation of the binding domain within U-gly showed that a 63 amino acid peptide retains the binding properties of U-gly and a 26 amino acid peptide within it (amino acids 695−720) is present in all U protein truncations competent to bind RNA. This region is absolutely required for the RNA binding activity of both the U protein and U-gly and most likely represents the minimal RNA binding domain. An interesting feature of this region is the presence of a cluster of RGG repeats. A search of the EMBL database

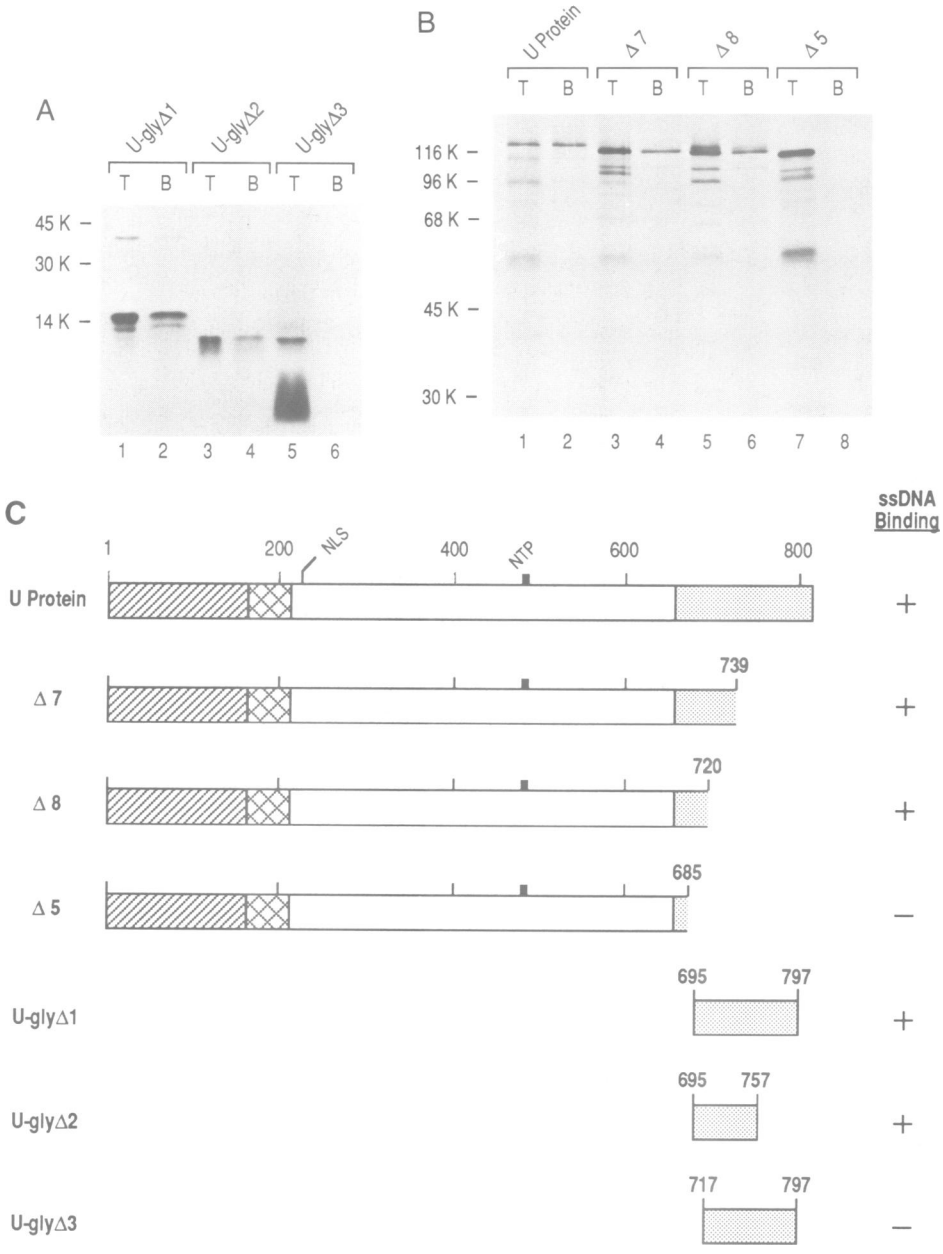

Fig. 7. The RGG box is an RNA binding domain. (**A**) Plasmids pGem695–797F, pGem695–757F and pGem717–797F were each linearized with *Pvu*II and used as template to generate *in vitro* transcripts with SP6 RNA polymerase and translated in wheat germ lysate to produce U-gly subdomains with amino acids 695–797 (U-glyΔ1; ~15 kDa), 695–757 (U-glyΔ2; ~10 kDa) and 717–797 (U-glyΔ3; ~10 kDa) respectively. All three peptides also contain 20 amino acids encoded by the vector polylinker at their C-terminus (see Materials and methods). Binding to ssDNA was carried out as described in the legend to Figure 3A. The 'T' lanes contain an aliquot of the total translation product used in the binding reaction shown in the 'B' lanes. (**B**) U protein and Δ5 were generated as described in the legend to Figure 3A. U protein having C-terminal truncations up to amino acid 739 (Δ7) and 720 (Δ8) were generated from plasmids pGemΔ3′739 and pGemΔ3′720 respectively and bound to ssDNA as described above. (**C**) The truncated proteins in (A) and (B) are represented schematically. The identity of the symbols are the same as in Figure 3B.

(Lipman and Pearson, 1985) for proteins that contain at least three RGGs revealed similar repeat structures in several other proteins (Figure 8). Interestingly, many of these are known RNA binding proteins and they contain RGG repeats at a characteristic spacing similar to that found in hnRNP U and usually interspersed with aromatic amino acids (Figure 8). We have termed this region the 'RGG box'.

RGG and RGGF peptides are found in many proteins including RNA binding proteins such as hnRNP K (Matunis *et al.*, 1992b). What is significant about the RGG box is that it contains several closely spaced RGG peptides. Our working definition of the RGG box is shown in Figure 8 which also suggests a consensus motif. We are currently uncertain of the minimal number of RGG (and GRG or RRG) repeats that are required for RNA binding activity.

Previous studies by several groups have pointed out the conservation of glycine-rich regions with interspersed arginine residues within RNA binding proteins and have suggested that these regions might be involved in binding RNA and/or in protein–protein interactions (Jong *et al.*, 1987; Christensen and Fuxa, 1988; Cobianchi *et al.*, 1988; Henriquez *et al.*, 1990; Suzuki *et al.*, 1991). By identifying the 26 amino acid region necessary for RNA binding we have demonstrated that the RGG box is required for RNA

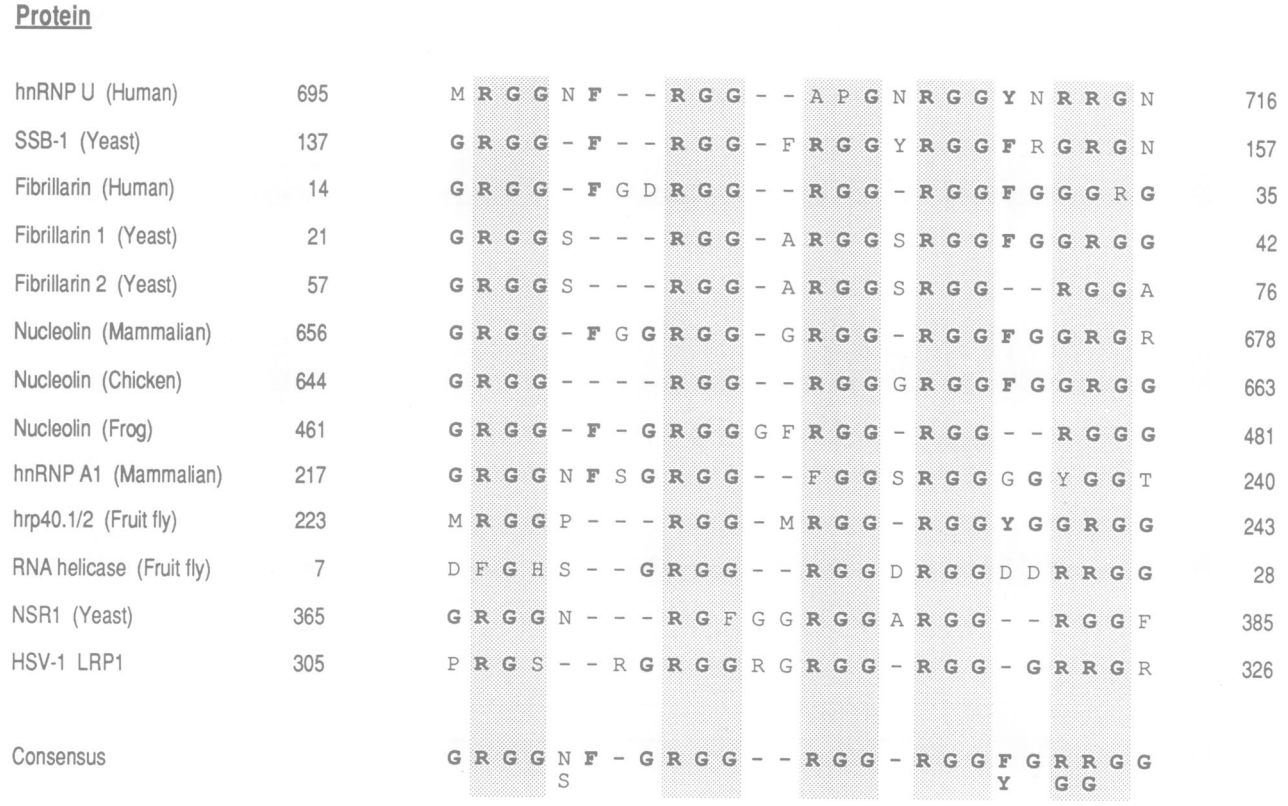

Fig. 8. Conservation of the RGG box in RNA binding proteins. The shaded area represents the most highly conserved RGG tripeptide. The bottom line is a consensus sequence and residues conforming to the consensus sequence are in bold. The first and last amino acid number within each sequence is indicated before and after the sequence respectively. Dashes are gaps placed in the sequence to obtain optimum alignment. Protein sequences were obtained from the following references: human hnRNP U (this paper); yeast SSB-1 (Jong *et al.*, 1987); fibrillarin from human (Aris and Blobel, 1991) and yeast (Henriquez *et al.*, 1990); nucleolin from human (Srivastava *et al.*, 1989), mouse (Bourbon *et al.*, 1988), hamster (amino acids 658–681, Lapeyre *et al.*, 1987), chicken (Maridor *et al.*, 1990) and frog (Caizergues-Ferrer *et al.*, 1989); hnRNP A1 from human (Buvoli *et al.*, 1988), and rat (Cobianchi *et al.*, 1986); fruitfly (*Drosophila melanogaster*) hrp40.1 and 2 (Matunis *et al.*, 1992a); RNA helicase from fruitfly (*D.melanogaster*, Dorer *et al.*, 1990); and yeast (*S.pombe* amino acids 506–528; *S.cerevisiae* amino acids 498–521, Iggo *et al.*, 1991); yeast NSR 1 (Lee *et al.*, 1991); and HSV-1 latency-related protein 1 (Wechsler *et al.*, 1989).

binding in the U protein and propose it to be an RNA binding motif. Aside from the RNA helicases (Dorer *et al.*, 1990; Iggo *et al.*, 1991), which would be expected to interact with RNA, and the herpes simplex virus-1 latency-related protein 1 (HSV-1 LRP1; Wechsler *et al.*, 1989) of unknown function, all of the RGG box-containing proteins listed have known RNA binding activity. The high degree of conservation of the RGG box in other RNA binding proteins suggests that it has the same function in the other proteins and that it is likely to be a predictor of RNA binding activity. An interesting example could be the HSV-1 LRP1 (Wechsler *et al.*, 1989). LRP1 is a putative protein predicted from the latency-associated transcript of HSV-1, however, it is uncertain if indeed it is produced during HSV infection. We would predict that if LRP1 is expressed, it would be a single-stranded DNA/RNA binding protein.

Many of the proteins listed in Figure 8 contain RBDs, and their binding to RNA is probably not mediated exclusively by the RGG box. However, the RGG box could potentially influence the overall binding property of a protein even if other binding domains are present as demonstrated for the hnRNP A1 protein. Cobianchi *et al.* (1988) have shown that the RBDs at the N-terminus of the hnRNP A1 protein require the glycine-rich carboxyl end for the cooperative binding of the protein to RNA. In the case of the U protein, the RGG box binding characteristics are sufficient for strong independent binding.

Although many different types of DNA binding motifs have been identified and characterized to date (Vinson *et al.*, 1989; Davis *et al.*, 1990; Pavletich and Pabo, 1991; Wolberger *et al.*, 1991; see Steitz, 1990 for review), relatively few RNA binding motifs have been identified (Dreyfuss *et al.*, 1988; Bandziulis *et al.*, 1989; Nagai *et al.*, 1990; Hoffman *et al.*, 1990; Calnan *et al.*, 1991b; Jessen *et al.*, 1991; see Steitz, 1990; and Kenan *et al.*, 1991, for reviews). One well characterized RNA binding domain, the RBD, is present in many RNA binding proteins (Dreyfuss *et al.*, 1988; Bandziulis *et al.*, 1989; Kenan *et al.*, 1991). Mutational analysis suggests that specific aromatic amino acids within the RBD are likely to interact with RNA and are essential for RNA binding (Scherly *et al.*, 1989; Lutz-Freyermuth *et al.*, 1990; Brennan and Platt, 1991; Jessen *et al.*, 1991). A second distinct RNA binding motif is the arginine-rich motif found in several viral, bacterial and ribosomal RNA binding proteins (Lazinski *et al.*, 1989). A short peptide cluster of arginine residues from the HIV-1 Tat protein can directly bind a specific RNA sequence and a single arginine in that cluster is responsible for the direct interaction (Calnan *et al.*, 1991b). Interestingly, U-gly does not contain either of these motifs. The U protein RGG box is particularly rich in glycine residues (~35%) but is also rich in arginine (~22%) and asparagine (~17%). Although there is a high arginine content, it does not have an arginine-rich cluster as seen in the typical arginine-rich RNA binding

domains. Similar to the DNA binding domain of many transcription factors, the RGG box is also very basic. However, the binding of U-gly (which contains the RGG box) to RNA is not simply due to a non-specific electrostatic interaction to the negatively charged phosphate backbone as it has a differential binding preference towards ribonucleotide homopolymers (Figure 6) and this binding is competed by ssDNA but not tRNA (M.Kiledjian and G.Dreyfuss, unpublished observations).

It is striking that the RGG boxes have strong positive charge (+3 to +9) but there are no lysines present. This strongly suggests that the arginine residues of the RGG box, and not simply the presence of positively charged residues, are required and are probably involved in RNA binding possibly in a similar fashion to that of HIV-1 Tat (Calnan *et al.*, 1991b). We also note the presence of aromatic residues in almost all RGG domains and these could contribute to hydrophobic stacking interactions with RNA bases. It is also interesting to note that the RGG box contains arginine residues flanked by glycines in the proximity of phenylalanines which are potential sites for dimethylarginine (DMA) modifications (Christensen and Fuxa, 1988). In fact several of the proteins listed in Figure 8 contain the modified residue DMA (Paik and Kim, 1989). Modification of the arginine residues may alter the RNA binding activity of these proteins which could provide a means to regulate their interaction with RNA.

Secondary structure predictions of the RGG box using the Chou–Fasman algorithm (Chou and Fasman, 1978) suggest that it is likely to form an unordered extended and flexible structure with turns at the RGGs. At present it is unclear how this domain can specifically interact with RNA. It is possible that the RGG box becomes ordered upon complexing with RNA as is believed to be the case with the arginine-rich binding domain of the HIV-1 Tat protein (Calnan *et al.*, 1991a). The glycines could provide multiple flexible hinges that allow the protein to conform to an ordered structure whereby the arginine(s) and/or aromatic amino acids come in contact with RNA. Understanding the mode of binding of the RGG box to RNA and of the effect of such binding on RNA structure will require experiments with synthetic RGG box peptides and mutagenesis of the protein.

Materials and methods

Isolation of cDNA clones and plasmid constructions

Mouse monoclonal antibody, 3G6 (Dreyfuss *et al.*, 1984b), was used at a dilution of 1:500 to immunoscreen a HeLa λZap II cDNA library (Stratagene) as previously described (Nakagawa *et al.*, 1986). One positive plaque was isolated and the ~1 kb *Eco*RI insert it contained was used to probe a HeLa D98 λgt11 cDNA library (kindly provided by Dr Tom Kadesch, University of Pennsylvania) by hybridization screening using standard techniques. One clone, U21.1 (described in Results), was inserted into pGem4 to generate pGem-U21.1. The entire U21.1 cDNA was sequenced on both strands by the dideoxy method (Sanger *et al.*, 1977).

Plasmids expressing N-terminal deletions of the U21.1 gene product, pGem-5′ΔApa, pGem-5′ΔBgl and pGem-5′ΔMsc, delete 5′ sequences up to the *Apa*I (nt 400), *Bgl*II (nt 1266) and *Msc*I (nt 2100) restriction endonuclease recognition sites, respectively. Plasmids expressing C-terminal deletions, pGemΔ3′Msc and pGemΔ3′Acc, remove 3′ sequences up to the *Msc*I (nt 2100) and *Acc*I (nt 1401) sites respectively. Plasmids pGemΔ3′739 and pGemΔ3′720 contain PCR-amplified sequences encoding U protein from amino acids 1–739 (nt 30–2258) and 1–720 (nt 30–2201), respectively flanked by *Eco*RI and *Bam*HI recognition sites. The plasmid pGemIΔSst removes the *Sst*I fragment of the U21.1 clone and expresses a protein with an internal deletion of amino acids 20–144 (nt 98–473).

Sequences encoding U protein amino acids 695–797 and 695–757 (pGem695–797F and pGem695–757F respectively) were amplified with PCR primers such that translation of these sequences extends into the pGem4 polylinker and generates 20 additional amino acids (DPLESTCRHAS-FRSPYSESY). The plasmid pGem717–797F was constructed similarly except that it expresses amino acids 717–797.

The entire coding region of the U protein was amplified by PCR using appropriate primers flanked by *Eco*RI sites, and was inserted into pMal-cRI (NEB) to generate pMal-U. This chimeric gene contains the U protein downstream of and in the same reading frame as the maltose binding protein for overexpression in bacteria. The plasmid pGem-MalU3′gly is the maltose binding protein with a eukaryotic consensus translation initiation site (Kozak, 1983) fused to the C-terminal amino acids 685–806 of the U protein inserted into pGem4 such that SP6 RNA polymerase could be used to generate an *in vitro* transcript. The plasmid containing the glycine-rich segment of the hnRNP A1 gene was derived from pHA-A1Gly (H.Siomi and G.Dreyfuss, manuscript in preparation) by excising the C-terminal 135 amino acids of A1 with *Eco*RI and *Pst*I and inserting it into pGem-4. All deletion and fusion constructs were confirmed by DNA sequencing. The yeast poly(A) binding protein-encoding plasmid pYEA3 has been described elsewhere (Adam *et al.*, 1986).

In vitro transcription/translation and immunoprecipitation

Plasmids were linearized at appropriate restriction sites 3′ of the desired translation stop codon to generate templates for *in vitro* RNA synthesis with SP6 polymerase and the resulting RNAs were translated in rabbit reticulocyte lysate or wheat germ extract in the presence of [^{35}S]methionine (Amersham) according to the manufacturer's suggested conditions (Promega Biotech).

Immunoprecipitations from HeLa cell nucleoplasm were carried out with the monoclonal antibody 3G6 bound to *Staphylococcus aureus* protein A in a 1% Empigen BB-containing buffer as previously described (Choi and Dreyfuss, 1984b).

Gel electrophoresis and immunoblotting

SDS–PAGE was performed as previously described (Dreyfuss *et al.*, 1984b). Immunoblotting was carried out with culture supernatant of the anti-hnRNP U monoclonal antibodies 8D3 and 11B8 (see below) or with 3G6, the anti-hnRNP U monoclonal antibody, ascites fluid diluted 1:1000 as described in Choi and Dreyfuss (1984b).

Ribonucleotide homopolymer and ssDNA binding assays

Binding of *in vitro* produced protein was carried out essentially as described in Swanson and Dreyfuss (1988) and Burd *et al.* (1991) with minor modifications. Briefly, ribonucleotide homopolymer (Pharmacia) and ssDNA–agarose (BRL) binding reactions were carried out with an equivalent of 10^5 counts per min (c.p.m.) of trichloroacetic acid (TCA)-precipitable protein in a total of 0.5 ml of binding buffer (Swanson and Dreyfuss, 1988) for 10 min on a rocking platform at 4°C . Unless otherwise stated, the NaCl concentration of the binding buffer was 250 mM. The beads were pelleted with a brief spin in a microfuge and washed five times with binding buffer prior to resuspension in 50 μl of SDS–PAGE loading buffer. Bound protein was eluted from the nucleic acid by boiling, resolved on a 12.5% SDS–PAGE gel and visualized by fluorography.

Production and purification of fusion protein

The pMal-U plasmid (see above) encoding the *Escherichia coli* maltose binding protein fused to the U protein (Mal-U) was expressed in *E.coli* TB1 cells and the fusion protein was partially purified on an amylose resin column as described by the manufacturer (NEB). Fractions containing Mal-U were dialyzed overnight at 4°C in 4 l of H_2O, lyophilized and resuspended in phosphate buffered saline.

Production of monoclonal antibodies

The U protein-specific monoclonal antibody 3G6 was prepared as previously described (Dreyfuss *et al.*, 1984b). Monoclonal antibodies 8D3 and 11B8 were raised against purified Mal-U fusion protein with intraperitoneal injections of 100 μg protein each into BALB/c mice as described in Dreyfuss *et al.* (1984b).

Peptide mapping

The HeLa U protein, U21.1 gene product and yeast poly(A) binding protein were cleaved at tryptophan residues with 2-(2′nitrophenylsulfenyl)-3-methyl-3-bromoindolenine (BNPS-Skatole, Pierce). U protein was immunopurified with 3G6 from [^{35}S]methionine-labeled HeLa cells (Piñol-Roma *et al.*, 1988) and run on SDS–PAGE side by side with *in vitro* translated U21.1 and yeast poly(A) binding protein. Proteins were transferred onto Immobilon-P Transfer Membrane (Millipore) according to the manufacturer's instructions and stained with Coomassie brilliant blue R (0.1% in 50%

methanol) for 5 min. Destaining was carried out with several rinses of 50% methanol/10% acetic acid followed by a rinse with 10% methanol prior to air drying and autoradiography. The appropriate bands were excised from the membrane and placed in 100% methanol to wet, and washed once in 50% methanol/10% acetic acid and twice in 70% acetic acid. Cleavage of the protein was performed directly on the membrane with 50 μl of 70% acetic acid containing 15 mM BNPS-Skatole (initially dissolved in anhydrous ether) at 42°C for 35 h in the dark (Omenn *et al.*, 1970). The BNPS-Skatole and acetic acid were removed with two 1 ml ether extractions. Protein was eluted as described by Szewczyk and Summers (1988) with 15 μl of elution buffer (50 mM Tris pH 8.8, 2% SDS and 1% Triton X-100) and microfuged for 10 min at room temperature. An equal volume of 2× SDS–PAGE sample buffer was added to the supernatant and the peptides were resolved by SDS–PAGE.

Acknowledgments

We thank Michael Matunis and Miriam Huizinga for help in generating monoclonal antibodies. We also thank members of our laboratory for critical reading of the manuscript. This research was supported by the Howard Hughes Medical Institute and grants from the National Institutes of Health.

References

Adam,S.A., Nakagawa,T.Y., Swanson,M.S., Woodruff,T. and Dreyfuss,G. (1986) *Mol. Cell. Biol.*, **6**, 2932–2943.
Aris,J.P. and Blobel,G. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 931–935.
Bandziulis,R.J., Swanson,M.S. and Dreyfuss,G. (1989) *Genes Dev.*, **3**, 431–437.
Bourbon,H.-M., Lapeyre,B. and Amalric,F. (1988) *J. Mol. Biol.*, **200**, 627–638.
Brennan,C.A. and Platt,T. (1991) *J. Biol. Chem.*, **266**, 17296–17305.
Burd,C.G., Swanson,M.S., Görlach,M. and Dreyfuss,G. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9788–9792.
Burd,C.G., Matunis,E.L. and Dreyfuss,G. (1991) *Mol. Cell. Biol.*, **7**, 3419–3424.
Buvoli,M., Biamonti,G., Tsoulfas,P., Bassi,M.T., Ghetti,A., Riva,S. and Morandi, C. (1988) *Nucleic Acids Res.*, **16**, 3751–3770.
Calnan,B.J., Biancalana,S., Hudson,D. and Frankel,A.D. (1991a) *Genes Dev.*, **5**, 201–210.
Calnan,B.J., Tidior,B., Biancalana,S., Hudson,D. and Frankel,A.D. (1991b) *Science*, **252**, 1167–1171.
Caizergues-Ferrer,M., Mariottini,P., Curie,C., Lapeyre,B., Gas,N., Amalric,F. and Amaldi,F. (1989) *Genes Dev.*, **3**, 324–333.
Christensen,M.E. and Fuxa,K.P. (1988) *Biochem. Biophys. Res. Commun.*, **155**, 1278–1283.
Choi,Y.D. and Dreyfuss,G. (1984a) *Proc. Natl. Acad. Sci. USA*, **81**, 7471–7475.
Choi,Y.D. and Dreyfuss,G. (1984b) *J. Cell Biol.*, **99**, 1997–2004.
Chou,P.Y. and Fasman,G.D. (1978) *Adv. Enzymol. Rel. Areas Mol. Biol.*, **47**, 45–148.
Cobianchi,F., SenGupta,D.N., Zmudzka,B.Z. and Wilson,S.H. (1986) *J. Biol. Chem.*, **261**, 3536–3543.
Cobianchi,F., Karpel,R.L., Williams,K.R., Notario,V. and Wilson,S.H. (1988) *J. Biol. Chem.*, **263**, 1063–1071.
Davis,R.L., Cheng,P.-F., Lassar,A.B. and Weintraub,H. (1990) *Cell*, **60**, 733–746.
Dorer,D.R., Christensen,A.C. and Johnson,D.H. (1990) *Nucleic Acids Res.*, **18**, 5489–5494.
Dreyfuss,G. (1986) *Annu. Rev. Cell Biol.*, **2**, 459–498.
Dreyfuss,G., Adam,S.A. and Choi,Y.D. (1984a) *Mol. Cell. Biol.*, **4**, 415–423.
Dreyfuss,G., Choi,Y.D. and Adam,S.A. (1984b) *Mol. Cell. Biol.*, **4**, 1104–1114.
Dreyfuss,G., Swanson,M.S. and Piñol-Roma,S. (1988) *Trends Biochem. Sci.*, **13**, 86–91.
Garcia-Bustos,J., Heitman,J. and Hall,M.N. (1991) *Biochim. Biophys. Acta*, **1071**, 83–101.
Henriquez,R., Blobel,G. and Aris,J.P. (1990) *J. Biol. Chem.*, **265**, 2209–2215.
Hoffman,D.W., Query,C.C., Golden,B.L., White,S.W. and Keene,J.D. (1990) *Proc. Natl. Acad. Sci. USA*, **88**, 2495–2499.
Hope,I.A. and Struhl,K. (1986) *Cell*, **46**, 885–894.
Iggo,R.D., Jamieson,D.J., MacNeill,S.A., Southgate,J., McPheat,J. and Lane,D.P. (1991) *Mol. Cell. Biol.*, **11**, 1326–1333.
Jessen,T.-H., Oubridge,C., Teo,C.H., Pritchard,C. and Nagai,K. (1991) *EMBO J.*, **10**, 3447–3456.
Jong,A.Y., Clark,M.W., Gilbert,M., Oehm,A. and Campbell,H. L. (1987) *Mol. Cell. Biol.*, **7**, 2947–2955.
Kay,B.K., Sawhney,R.K. and Wilson,S.H. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1367–1371.
Kenan,D.J., Query,C.C. and Keene,J.D. (1991) *Trends Biochem. Sci.*, **16**, 214–220.
Kozak,M. (1983) *Microbiol. Rev.*, **47**, 1–45.
Lapeyre,B., Bourbon,H. and Amalric,F. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 1472–1476.
Lazinski,D., Grzadzielska,E. and Das,A. (1989) *Cell*, **59**, 207–218.
Lee,W-C., Xue,Z. and Melese,T. (1991) *J. Cell Biol.*, **113**, 1–12.
Lipman,D.J. and Pearson,W.R. (1985) *Science*, **227**, 1435–1441.
Lutz-Freyermuth,C., Query,C.C. and Keene,J.D. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6393–6397.
Maridor,G., Krek,W. and Nigg,E.A. (1990) *Biochim. Biophys. Acta*, **1049**, 126–133.
Matunis,E.L., Matunis,M.J. and Dreyfuss,G. (1992a) *J. Cell Biol.*, **116**, 257–269.
Matunis,M.J., Michael,W.M. and Dreyfuss,G. (1992b) *Mol. Cell. Biol.*, **12**, 164–171.
Nadler,S.G., Merrill,B.M., Roberts,W.J., Keating,K.M., Lisbin,M.J., Barnett,S.F., Wilson,S.H. and Williams,K.R. (1991) *Biochemistry*, **30**, 2968–2975.
Nagai,K., Oubridge,C., Jessen,T.H. and Evans,P.R. (1990) *Nature*, **348**, 515–520.
Nakagawa,T.Y., Swanson,M.S., Wold,B.J. and Dreyfuss,G. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 2007–2011.
Omenn,G.S., Fontana,A. and Anfinsen,C.B. (1970) *J. Biol. Chem.*, **245**, 1895–1902.
Paik,W.K. and Kim,S. (1989) *Protein Methylation*. CRC Press, Inc., Boca Raton, FL, pp. 98–123.
Pavletich,N.P. and Pabo,C.O. (1991) *Science*, **252**, 809–817.
Piñol-Roma,S., Choi,Y.D., Matunis,M.J. and Dreyfuss,G. (1988) *Genes Dev.*, **2**, 215–227.
Rossmann,M.R., Moras,D. and Olsen,K. (1974) *Nature*, **250**, 194–199.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Scherly,D., Boelens,W., Dathan,N.A., van Venrooij,W.J. and Mattaj,I.W. (1989) *EMBO J.*, **8**, 4163–4170.
Srivastava,M., Fleming,P.J., Pollard,H.B. and Burns,A.L. (1989) *FEBS Lett.*, **250**, 99–105.
Steitz,T.A. (1990) *Q. Rev. Biophys.*, **23**, 205–280.
Suzuki,K., Olvera,J. and Wool,I.G. (1991) *J. Biol. Chem.*, **266**, 20007–20010.
Swanson,M.S. and Dreyfuss,G. (1988) *Mol. Cell. Biol.*, **8**, 2237–2241.
Swanson,M.S., Nakagawa,T.Y., LeVan,K. and Dreyfuss,G. (1987) *Mol. Cell. Biol.*, **7**, 1731–1739.
Szewczyk,B. and Summers,D.F. (1988) *Anal. Biochem.*, **168**, 48–53.
Vinson,C.R., Sigler,P.B. and McKnight,S.L. (1989) *Science*, **246**, 911–916.
Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) *EMBO J.*, **1**, 945–951.
Wechsler,S.L., Nesburn,A.B., Zwaagstra,J. and Ghiasi,H. (1989) *Virology*, **168**, 168–172.
Wolberger,C., Vershon,A.K., Liu,B., Johnson,A.D. and Pabo,C.O. (1991) *Cell*, **67**, 517–528.

Received on March 11, 1992; revised on April 9, 1992

Note added in proof

A recent report from Girard *et al.* [(1992) *EMBO J.*, **11**, 673–682] describes a new glycine–arginine-rich protein from yeast that affects the synthesis of 18S ribosomal RNA. We note that this protein contains three RGG boxes. Furthermore, a recent paper from Ghisolfi *et al.* [(1992) *J. Biol. Chem.*, **267**, 2955–2959] demonstrates that the region of nucleolin, which contains an RGG box, interacts with RNA directly and adopts multiple β-turn structures. The nucleotide sequence of U21.1 cDNA reported in this paper will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under the accession number X65488.