Nucleic Acids Research, Vol. 19, No. 18 ***4931–4936***

U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts

Donald E.Tsai, David S.Harper and Jack D.Keene*
Department of Microbiology and Immunology, Duke University Medical Center, Durham, NC 27710, USA

Received June 13, 1991; Revised and Accepted August 21, 1991

ABSTRACT

The U1snRNP-A (U1-A) protein was used to select specific RNA sequences from a degenerate pool of transcripts using direct RNA binding and polymerase chain reaction amplification (PCR). Sequences were randomized in loops of 10 or 13 nucleotides or as a linear stretch of 25 nucleotides. From all three structural contexts, an unpaired ten nucleotide consensus sequence was obtained. A selected stem-loop structure that resembled the natural U1-A protein binding site on loop II of U1 RNA demonstrated the highest affinity of binding in comparison with the other structural contexts. A data profile of selected sequences identified U1 RNA upon searching the GenBank database. Thus, this method was useful in determining the sequence specificity of an RNA binding protein and may complement the use of phylogenetic comparisons to predict conserved recognition elements. These findings also suggest that the evolutionary conservation of loop II of U1 RNA results from constraints imposed by protein binding.

INTRODUCTION

A common RNA recognition motif (RRM) consisting of approximately 80 amino acid residues is shared by a family of proteins, and constitutes part of a specific RNA-binding domain in the cases of poly A-binding (PAB), U1 snRNP-70K, U1-A, and U2 snRNP-B″ proteins (U2-B″) (1–7). Members of this family bind a variety of RNA molecules such as oligouridylates 3′ to stem-loop structures (La protein, (8)), snRNA stem-loop structures (U1snRNP-70K, (2), U1-A, (5,6), and U2-B″, (7,9) proteins), and adenylate homopolymers (PAB protein, (1)). The tertiary structure of RRM 1 of the U1-A protein has recently been determined and other RRM proteins have been modeled based upon the U1-A protein structure (10–12). RNA recognition in these cases appears to involve conserved structural features in the RRM while binding specificity is dependent on nonconserved amino acid residues (6,13). While the number of identified RRM-containing proteins continues to increase, the specific RNA ligand in the majority of cases remains unknown.

We have previously selected specific RNA species from a pool of total HeLa cell RNA based on their ability to bind to RRM-containing proteins (4,5,7) or to RNA-specific autoantibodies (14). These methods allow coimmunoprecipitation of U1 or U2 RNAs, produced either in vitro or in vivo, with their cognate binding proteins. However, most cellular RNAs are much less abundant than U1 or U2 RNAs, and may not be detectable by standard coimmunoprecipitation assays. Selection from a pool of degenerate nucleic acids (random RNA selection) has been used to identify the sequence specificity of other nucleic acid binding proteins (15,16). We have used the U1-A protein to determine whether this type of selection is useful for determining the RNA ligand of an RRM containing protein. Selection from a pool of degenerate RNAs should also provide more information on sequence constraints imposed upon U1 RNA as required for U1-A protein binding.

The U1-A protein is an RRM family member in which one of two RRMs binds specifically to stem-loop II of U1 RNA (6,17). A similar sequence is present in U2 snRNA stem loop IV, which has a loop of 13 nucleotides and binds specifically to the U2-snRNP B″ (U2-B″) (7,18), but very weakly to U1-A protein (7). In the present study, in vitro RNA transcripts containing degenerate nucleotide sequences in both stem-loop and linear contexts were generated and selected for binding to U1-A protein. Selected RNA species were reverse transcribed, amplified by the polymerase chain reaction (PCR) (19), cloned and sequenced. A set of selected RNAs with a consensus of 10 nucleotides (nts) was derived that matched the natural loop of stem-loop II of U1 RNA (5,6). We compared this approach with phylogenetic analysis as a means of identifying sequences important for protein binding to RNA ligands. We interpret these findings with respect to the elements of RNA recognition between the U1-A protein and U1 RNA.

MATERIALS AND METHODS

Enzymes and biochemicals

Restriction enzymes, T7 RNA polymerase, T4 DNA ligase, Sequenase version 2.0 DNA polymerase and AMV reverse transcriptase were obtained from United States Biochemical.

* To whom correspondence should be addressed

RNasin and pSp64 were obtained from Promega. Taq polymerase was purchased from Stratagene. Nucleotides were obtained from Pharmacia. [α-^{32}P] UTP was obtained from ICN Biomedicals. *E.coli* carrier tRNA and protein-A Sepharose beads were purchased from Sigma. The oligodeoxyribonucleotides were synthesized using an Applied Biosystems DNA synthesizer.

Preparation of gene 10 fusion protein

RNA-binding reactions contained U1-A protein derived polypeptides linked at the amino terminus to the first 12 amino acids of phage T7 gene 10 (g10) protein, which serves as an antigenic tag, produced in pET-3c vectors as described previously (20). *Escherichia coli* strain BL21(DE3) was used to prepare protein extracts of the above fusion proteins as described (7,20). Rabbit antiserum specific for the gene 10 epitope was provided by Daniel Kenan.

Preparation of RNA

Double-stranded transcription template was prepared by subjecting 5 ng of Stem-Loop N10, Stem-Loop N13 or Linear N25 oligodeoxynucleotide (see figure 1 for description of oligonucleotides) to 35 cycles in an Ericomp Temperature Cycler (1 min. 94°C, 1 min. 50°C, 2 min. 72°C) in the following buffer: 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin, 0.1 μg primer T7Univ, 0.1 μg primer RevUniv, 200 uM dNTP and 2.5 units Taq DNA polymerase (Fig. 1). Any tandem products were reduced to monomers by cutting with BamHI. The DNA template was then transcribed using T7 RNA polymerase (21).

Selection of RNA by coimmunoprecipitation

1−5 μg of full length g10 U1snRNP-A fusion proteins was bound to 4 mg of protein-A beads using the antibody against the gene 10 epitope. After three washes with NT2 buffer (4), the protein and 200−500 ng of RNA were incubated for 7 minutes in 100 μl RNA binding buffer (4). Following this incubation, the RNA was washed 5 times with NT2 buffer. 10 μg of carrier tRNA was then added to the immunoprecipitated RNA. It was then subjected to phenol extraction and ethanol precipitation. Rounds two and three had identical binding conditions except that 0.5 M urea was added to the final NT2 buffer wash to reduce nonspecific binding as described previously (4).

Reverse transcription and reconstitution of the transcription template

One third of the bound

RNA was reverse transcribed using 0.1 μg primer RevUniv with AMV reverse transcriptase (1 hour at 42° using conditions recommended by the supplier). Following reverse transcription, the cDNA was resuspended in 10 μl double distilled water. 3 μl of the cDNA was subjected to 35 cycles of PCR under the conditions already noted. The new transcription template was then used to repeat the above transcription, coimmunoprecipitation and reverse transcription steps for two additional rounds as outlined in figure 2.

Cloning and sequencing

After 3 total rounds as specified above, the double stranded BamHI cut transcription template was cloned into pSp64. The recombinants were used to transform HB101 cells which were grown overnight on ampicillin LB agar plates. Plasmid DNA was prepared from the transformed cells and sequencing was performed by the dideoxy chain termination method using United States Biochemical Sequenase Version 2.0.

Computer analysis

All sequences obtained were combined to make a sequence profile using the Wisconsin Software 'profile' command (22). This sequence profile was then used to search the eukaryotic library in Genbank to identify similar sequences.

Quantitative electrophoretic mobility shifts

A constant amount of ^{35}S-labeled in vitro translated U1-A protein, amino acids 11 to 96 was incubated as described (4) with varying concentrations of in vitro transcribed U1, A1, B7, or D2 RNAs each at varying concentrations. Complexes were separated on nondenaturing 5.0% polyacrylamide/90 mM Tris Borate (pH 8.3) gels (23). Gels were soaked in 25 mM 1.3-bis[tris(hydoxymethyl) methylamino]propane (pH 10.0) in 0.1 N NaOH for 10 minutes to deacylate [^{35}S]methionyl tRNA, soaked in 0.5 M sodium salicylate for 20 minutes, dried, and fluorographed. Quantitative analysis to determine relative binding affinities was done as described (5,7,23).

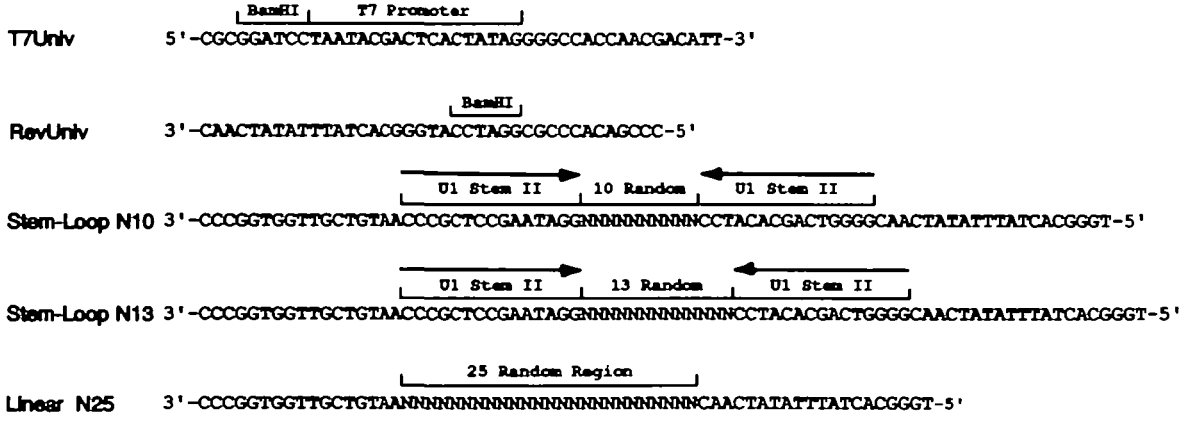

Figure 1. Oligonucleotides used in the reverse transcription/PCR selection process to yield different structural contexts. Nucleotide sequences of synthetic DNA oligonucleotides used in this study are shown. Horizontal arrows indicate the complementary regions of the U1 stem. Restriction sites, T7 promoter, U1 stem II, and degenerate regions are noted by lines above the specific nucleotides. Degenerate nucleotides are represented by 'N'.

RESULTS

RNA binding specificity of U1 snRNP-A protein

The isolation of specific RNA species from pools of total cell RNA using the g10 epitope tag on various RNA binding proteins was described previously (5,7). In the present study, we created a degenerate pool of RNA using synthetic DNA oligomers that were randomized in three different structural contexts (Fig.1) in order to study the preferred binding properties of U1-A protein. To assure a thorough selection, multiple rounds of transcription, RNA binding, coimmunoprecipitation, reverse transcription, and PCR were carried out before vector cloning as outlined in figure 2. The structural contexts used in this study were the stem of stem-loop II of U1 RNA (24) with a degenerate loop of 10 or 13 nts and a linear RNA of 25 nts which were all degenerate (Fig. 1).

Sequence analysis of multiple clones, each representing a distinct coimmunoprecipitated RNA species, revealed a recognition consensus sequence (RCS) favored by U1-A protein, that was present in all four context classes (A ,B ,C, and D)(Fig.3 shaded). A fourth context, designated class C, was inadventantly created by a PCR mutation (25). The RCS can be divided into two regions, one with a high level of conservation and the other with more variability (Fig. 4). The highly conserved region is represented by the 5′ seven nucleotides (AUUGCAC), 87.5% to 100% conserved at each position. The same sequence was present in the context of linear RNA and hairpin stem-loops with possible loop sizes of 8, 10, 11, 13 and 15 nucleotides (Fig. 3). In some cases, marked by an asterisk, the loop size may be smaller due to potential base pairing between the 5′ and 3′ most nucleotides in the loop, thus possibly increasing the stem length and creating loop sizes of 8 and 11 nucleotides. Following the 7 highly conserved nucleotide positions, there is a much weaker preference in the next three nucleotides. A detectable preference in the 3-mer region favors the wild type human U1 RNA sequence in positions 8, 9, and 10.

The context of the RCS did not appear to be restrictive for binding as RNAs from all three contexts were selected. RNAs from class A which have a loop of 10 nucleotides displayed no flexibility in nucleotide position as the loop size was identical to the native recognition site (5,6). All 9 species from this group were aligned identically with no variability in the position of the conserved sequences within the loop and all 10 nucleotides of the loop were involved (Fig. 3). RNAs from class B and C with potential loop sizes of 11 (members of class B with asterisk, Fig. 3), 13, or 15 nucleotides had flexibility in the position of the 10 nucleotide binding site with the extra nucleotides both 5′ and 3′ to the RCS (Fig 3). Linear RNA with no stem structure contained the RCS at random positions. While it may appear that the RCS is biased towards the 5′ end of the degenerate region, this is not the case. The RCS of RNAs D1-D5 begins at the terminal 3 nucleotides of the T7Univ primer ('ATT' Fig. 1), statistically favoring the selection of RNA with the RCS at the very 5′ end of the degenerate region (Fig. 3). Thus, the bias was attributable to the primer design rather than a selective preference of the protein. Some linear clones were selected which had less than the original 25 nucleotides in the degenerate region possibly due to artifacts generated by the PCR.

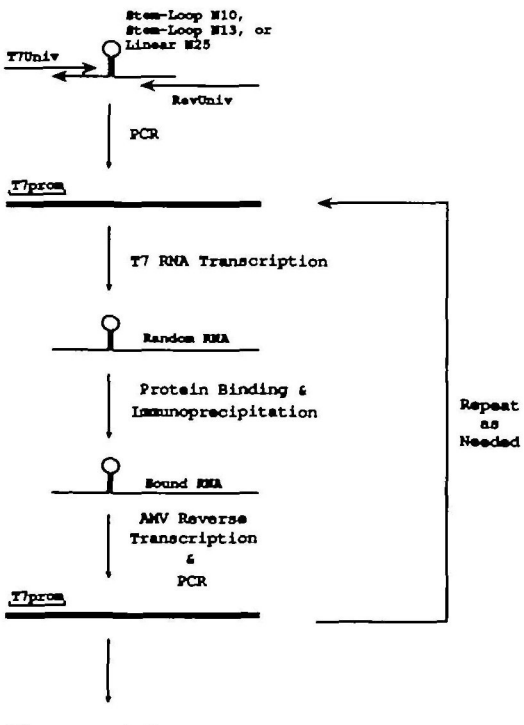

Figure 2. The RNA selection protocol. The sequential steps of selection and amplification are detailed in the Materials and Methods. The T7 promoter region is marked by *T7prom*. DNA and RNA are represented by double lines and single lines, respectively.

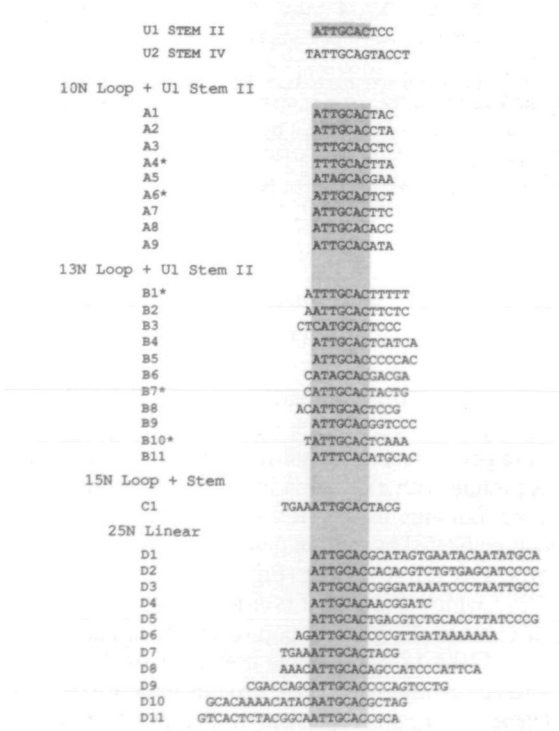

Figure 3. Alignment of the four classes of selected RNA sequences. The specific sequences corresponding to the originally degenerate regions are shown for each different structural context in which RNA was selected (classes A, B, C, and D). The conserved 7mer region of the RCS has been aligned and shaded in both the selected sequences and wild type human U1 loop II. Asterisks represent the possibility that the sequence of the loop may form an additional base pair between the 5′ and 3′ nucleotides. C1 was probably formed by a PCR induced mutation changing the 'C' to a 'T' in the last base pair of the stem in a 13 nucleotide loop.

Position	1	2	3	4	5	6	7	8	9	10
A	28 (87.5%)	1 (3.1%)	2 (6.2%)	0 (0.0%)	0 (0.0%)	32 (100.0%)	0 (0.0%)	5 (15.6%)	8 (25.0%)	9 (28.1%)
C	1 (3.1%)	0 (0.0%)	0 (0.0%)	1 (3.1%)	32 (100.0%)	0 (0.0%)	32 (100.0%)	7 (21.9%)	12 (37.5%)	17 (53.1%)
G	0 (0.0%)	0 (0.0%)	0 (0.0%)	31 (96.9)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (18.7%)	4 (12.5%)	2 (6.3%)
T	3 (9.4%)	31 (96.9%)	30 (93.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	14 (43.8%)	8 (25.0%)	4 (12.5%)
RCS:	A	T	T	G	C	A	C	var	var	var

Nucleotide

Figure 4. Compilation of the selected RNA sequences and the Recognition Consensus Sequence (RCS) for U1-A protein. The ten nucleotides present in each selected RNA are aligned and the total number and percentage of each nucleotide in the ten positions was calculated. The sequence of the RCS is shown at the bottom. Nucleotides in the highly conserved 7mer of the RCS are boldfaced. The 3 variable nucleotide positions of the RCS are marked 'var'.

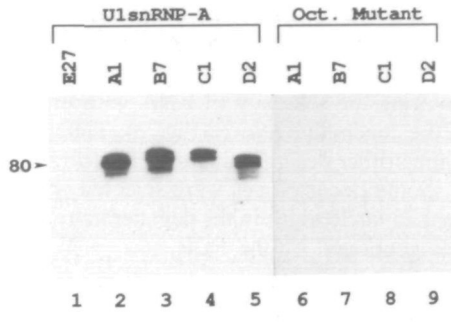

Figure 5. Immunoprecipitation of radiolabeled RNA by full length wild-type U1-A protein and first RRM octamer mutant of U1-A protein. In vitro transcribed ^{32}P-labeled RNA was coimmunoprecipitated by the anti-g10 antiserum along with either wild type g10 U1-A or g10 U1-A with a double RNP1 octamer mutation. The bound RNA was analyzed on an 8% acrylamide gel and exposed to film. The sequences tested are representative of each of the different classes. E27 represents a non-RCS loop in the stem-loop N10 context as a negative control.

Comparative binding properties of selected RNAs using the wild type and a mutant in the RNA binding domain of U1-A protein

Individual coimmunoprecipitations of ^{32}P-labeled RNA transcripts representive of each class (A1, B7, C1, and D2, Fig. 3) were performed using full-length wild type U1-A protein and U1-A protein with a mutation in the first RRM that is known to eliminate binding to U1 RNA (17). These RNAs were coimmunoprecipitated by the anti-gene10 antibody in the presence of wild-type g10 U1-A protein (Fig 5, lanes 2−5). A negative control E27, which contains a non-RCS loop sequence in the context of U1 stem II, was not coimmunoprecipitable by U1-A protein (Fig.5 lane 1). None of the transcripts were precipitated in the presence of the g10 U1-A protein mutant (Fig. 5 lanes 6−9). These results are consistent with previous findings that RNA binding can be attributed to the first RRM of U1-A protein (6,17).

Relative binding affinities of selected RNAs

The relative binding affinities of U1-A protein for class A, B, and D RNAs containing the RCS were determined. Representative clones from each group were transcribed in vitro and incubated with the ^{35}S labeled, in vitro-translated amino-terminal RNA binding domain of U1-A protein. Various concentrations of unlabeled RNA were incubated with a constant amount of labeled protein in order to determine relative affinities, as described previously (7,17). After incubation, samples were analyzed on 5% native acrylamide gels (Fig. 6). The nanomolar amount of RNA, and the species of RNA used in each lane are indicated. The results demonstrated that both the wild type U1 RNA and A1 RNA have similar affinities for U1-A protein (see lanes 3−7 for A1 RNA and 8−12 for wild type U1 RNA). The affinities of B7 RNA (lanes 13−16) and linear D2 RNA (lanes 17−20) for the U1-A protein were approximately 100 fold lower based on densitometric analysis of the autoradiographs (not shown). These data suggest that while the RCS is sufficient to allow interaction between these RNAs and U1-A protein, the natural RNA conformation of U1 stem-loop II with a 10-mer loop that is common to both the wild type U1 snRNA and to class A sequences is preferred.

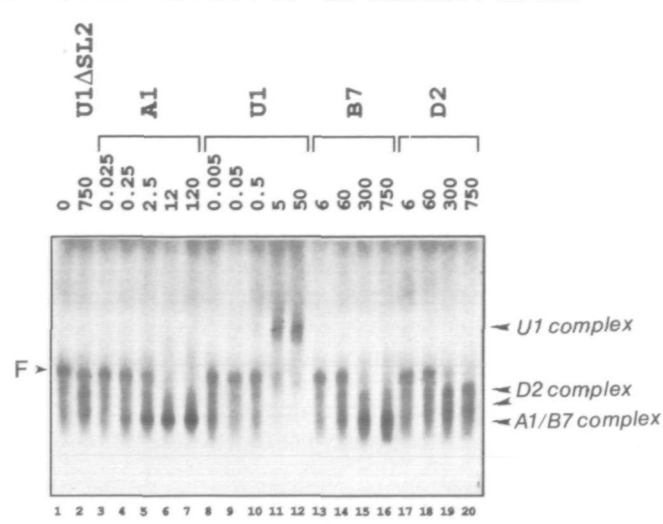

Figure 6. Quantitative electrophoretic mobility shifts of representative sequences from each class of RNA. ^{35}S-labeled in vitro translation product representing the first RRM (amino acids 11−96) of U1-A protein was incubated with unlabeled RNA and assayed for complex formation on a 5% nondenaturing polyacrylamide gel and exposed to film: Lane 1, no additional RNA (tRNA competitor only), lane 2, deletion mutant of U1RNA lacking stem loop II; lanes 3−7, increasing concentrations of RNA-A1; lane 8−12, increasing concentrations of wild type U1 RNA; lanes 13−16, increasing concentrations of RNA-B7; lanes 17−20, increasing concentrations of RNA-D2. Concentrations (in nM) of each unlabeled RNA are indicated at the top of each lane. Unbound U1-A is represented by 'F'.

Would the natural cognate RNA be identified from this random selection procedure?

The sequences derived using the 10 nucleotide loop, the 13 nucleotide loop and the 25 nucleotide linear RNAs were combined and transformed into a sequence profile using the Wisconsin Software 'profile' function (22). A search of eukaryotic sequences in Genbank using this profile resulted in the 10 best matching sequences (Table 1). It should be noted that the top score was from a human U1 snRNA which is the correct cognate RNA. Seven out of the top 10 sequences are various forms of U1 snRNA from different organisms. This implies that by using a sequence profile derived from a large group of sequences, one can identify the natural cognate RNA, if its sequence has been reported. Whether this in vitro RNA selection approach will have more general application to other RNA-binding proteins is to be determined.

Table 1. Sequences obtained from a GenBank profile search that most closely match the RCS of U1-A Protein. The ten sequences are listed top to bottom in order of best match.

Human U1A snRNA
Calf U1A snRNA
Human thyroxine binding globulin
Chicken non-functional alpha-2(I) collagen mRNA
Rat U1 snRNA
Mouse U1b-2 snRNA
Human U1 snRNA candidate gene 3.84
Human U1.11 snRNA pseudogene
Artemia salina satellite (negative strand)
Human U1 snRNA gene HSD5 12/89

DISCUSSION

Identification of an RCS

RRM-containing proteins constitute a superfamily whose known members have increased rapidly. However, determination of the binding specificities and cellular functions of new family members has not proceeded as quickly. Selection of RNAs that bind to a protein using a degenerate oligonucleotide pool may be useful in identifying the cognate RNA ligand. Using full length U1-A protein linked to the g10 epitope tag, we coimmunoprecipitated highly degenerate in vitro RNA transcripts derived from three template classes. Class A represented the framework for the natural U1 stem loop II, having the identical stem but with a degenerate 10-mer loop of 1.0×10^6 degeneracy. Class B represented a hybrid between U1 and U2 snRNAs with a stem from U1 RNA stem-loop II, but with a 13-mer loop similar to that in U2 stem-loop IV with 6.7×10^7 degeneracy. Class D was derived from a linear stretch of 25 nucleotides of 1.1×10^{15} degeneracy but with no preset context other than the flanking PCR primer sites. Class C was assumed to have arisen from a PCR-induced mutation in the stem-loop of a Class B clone, giving rise to an RNA loop size of 15 nucleotides. All classes of RNA contained identical fixed flanking sequences designed to provide efficient PCR primer sites without forming potential secondary structures with other regions of the oligonucleotide. Five ng of each oligonucleotide were used in the first PCR reaction to create the original degenerate templates. Therefore, some species of the linear class D construct were not presented for binding because this amount of oligonucleotide provides approximately 1.52×10^{11} species. On the other hand, for the 10mer and 13mer stem-loop constructs this statistical bias was not a concern because 1.22×10^{11} and 1.17×10^{11} species of RNA were presented respectively. Cloning the final PCR products and sequencing individual clones gave rise to a collection of sequences consistently representing a portion of the RNA binding site of U1-A as determined previously (5,6). Examination of the sequences showed that an RCS was evident that comprised a seven nucleotide region that is highly conserved at each position, followed by a three nucleotide region of greater variability.

All RNA sequences obtained formed a single RCS. This RCS was shown to bind only the amino-terminal RRM but not the full length U1-A protein that contained a mutation disrupting the amino-terminal RRM (Fig. 5). The RNA binding ability of the carboxyl-terminal RRM of U1-A protein was not evident in our study. This was possibly due to the absence of additional factors necessary for RNA binding (18,26), the use of stringency levels that did not permit the carboxyl-terminal RRM to bind its cognate RNA or the possibility that the carboxyl-terminal RRM may not be capable of binding to any RNA. A final possibility is that the required RNA was not present in the original degenerate RNA pool either due to statistical omission from the original pool or due to preference for an RNA binding site in excess of 25 nucleotide.

As a stem-loop structure confers higher affinity binding versus a linear RNA (Fig. 6), one might have expected to find stem-loop RNAs among the class D group. The absence of these RNAs could be due to the stringency conditions used in the coimmunoprecipitation step. We used conditions for binding and washing that have been previously shown to allow specific binding of U1snRNP-70K, U1snRNP-A and U2snRNP-B″ to their respective natural cognate RNA with minimal detectable background from a pool of total HeLa cell RNA (4,5,7). In the case of the class D linear RNAs, it is possible that no stem-loop structures were detected because the conditions may have been of low enough stringency to allow the more numerous linear RNAs to bind as effectively as any stem-loop RNA formed, thus statistically masking the latter group. By using a higher level of stringency it may have been possible to filter out the linear RNAs from stem-loop RNAs. It is likely that conditions will have to be optimized for each specific case.

Structural context

The context of the RCS was not critical for recognition, in that all three classes selected the same RCS using the coimmunoprecipitation method. Interestingly, due to the poor conservation in the last three nucleotides, the sequence for wild type U1 loop II was not found among the samples that were sequenced. While the loop context was not critical for recognition during the selection process, the mobility shift data indicated that the binding affinity was affected by the structural context. For example, the U1-A protein binding affinities of both the B7 RNA and the linear D2 RNA were approximately 100-fold less than that of A1 RNA or wild type U1 RNA. In the case of Class D, we did not select any stem-loop structure similar to the native U1-A protein binding site on U1 RNA when using the linear sequence 25N. Thus, the RNA selection method applied here was limited in its ability to select the binding sequence in its natural context when using a random linear sequence.

Our results correlate well with previous findings of the U1 RNA binding specificity of U1-A and U2-B″ proteins as shown by mutational analysis (9). It was shown that a cytidine in position 7 in the U1 snRNA loop II is critical for binding by U1-A protein while a guanosine would confer U2-B″ binding. In addition, an adenosine inserted into the loop of U1 RNA at position 9 seemed to favor U2-B″ binding. Thus, our data (Fig. 4) are consistent with the conclusion that position 7 is critical because one hundred percent of our sequences in all three contexts contained a cytidine in this position. This is consistent with the extremely weak binding of U1-A protein to U2 RNA, which has a very similar sequence in loop IV but contains a guanosine at the corresponding position 7 (7,9). The data suggests that the last three nucleotides are not critical for binding. An adenosine inserted in position 9 of the U1 RNA loop in the study by Scherly et al.(9) decreased binding by U1-A protein. By our selection data, it appears that the RCS (Figs 3 and 4) can easily tolerate nucleotide changes in this position. In addition, clone A1, which contains the 'U' to 'A' change in position 9, has approximately the same binding strength as wild type U1 RNA as determined by our mobility shift assays (Fig. 6). Thus, the disruption of binding observed by Scherly et al. probably resulted from the change in the loop size rather than from the nucleotide substitution.

The B7 and D2 RNAs bind to U1-A protein with similar affinities, but about 100 fold less than the A1 RNA. These findings are consistent with work on the U1 RNA variants found in the mouse. Two major U1 RNA variants, mU1a and mU1b, are differentially expressed during development with mU1a present mostly in differentiated cells and mU1b mostly in stem cells (27). In mU1a, the adenosine in position 6 of stem-loop 2 is methylated, positions 9 and 10 contain cytidines, and the loop size is 10. In mU1b, the adenosine in position 6 is unmethylated, positions 9 and 10 contain thymidines and the stem is altered so that the loop size is 11. These differences have been suggested to lower the affinity of the U1-A protein for mU1b compared to mU1a (28). This is consistent with our findings that alteration of the context of the RCS can drastically alter the affinity of the binding by U1-A protein. The functional relevance of differential U1-A protein binding to variants of U1 RNA and other RNAs containing the RCS has not been determined.

Comparison of phylogenetic conservation

The sequence specificity of the RCS correlates well with phylogenetic analysis of U1 RNA sequences (29). The strongly conserved 7 mer region of the RCS was found to be highly conserved by phylogenetic comparison. Positions 9 and 10, which are in the variable 3 mer region of the RCS, are the most poorly conserved positions among species. Evolutionary conservation of loop sequences has been interpreted to suggest that these nucleotides are critical for protein binding. Our data demonstrate that the conservation of the loop II sequence of U1 RNA can be accounted for by the constraints imposed by binding of the U1-A protein. While there is agreement between results obtained by our selection procedure and phylogenetic comparisons, a critical difference occurs in position 8. This position is strongly conserved across species, yet was relatively variable in our studies. This discrepancy opens the possibility that nucleotide 8 may have been conserved for functions independent of U1-A protein binding.

Relationship of RCS to the native cognate RNA

A major problem with studying RRM-containing proteins has been the difficulty of identifying the cognate RNA. A potential application of the random selection approach is that one might be able to identify the in vivo cognate RNA sequence by selection from a pool of degenerate RNAs in vitro. In work by Tuerk and Gold using T4 DNA polymerase, only two products were present following three rounds of selection by filter binding, one of which was the known cognate RNA. Our work using U1-A protein yielded a pool of at least 31 different RNAs each containing a common core sequence. It is possible that the T4 polymerase is more specific in its binding requirements than U1-A protein. Alternatively, the stringency of the binding assay may have precluded T4 DNA polymerase from binding RNAs at low affinity. In the case of U1-A protein, there seem to be a number of different RNA sequences that can bind under our conditions. If the wild type sequence had a substantially higher affinity than the other sequences in the pool, one would expect it to out compete the others after three successive cycles. No wild type was found, however, indicating that it has a comparable affinity but was not selected for statistical reasons.

If U1 RNA had not been known as the cognate RNA for U1-A protein, our method of selection and computer analysis would have identified it. But given that U1 RNA is abundant in most eucaryotic cells and may be over represented in the database, U1-A protein may be an idealized example. Thus, this approach may not be applicable to all proteins with unknown binding specificities. However, once the cognate RNA is in a database, this procedure maybe useful in identifiying it and for discerning critical nucleotide requirements for binding.

REFERENCES

1. Sachs, A.B., Davis, R.W. and Kornberg, R.D. (1987) *Mol. Cell. Biol.*, **7**, 3268–3276.
2. Query, C.C., Bentley, R.C. and Keene, J.D. (1989) *Mol. Cell. Biol.*, **9**, 4872–4881.
3. Bandziulis, R.J., Swanson, M.S. and Dreyfuss, G. (1989) *Genes Dev.*, 3, 431–437.
4. Query, C.C., Bentley, R.C. and Keene, J.D. (1989) *Cell*, 57, 89–101.
5. Lutz-Freyermuth, C. and Keene, J.D. (1989) *Mol. Cell. Biol.*, **9**, 2975–2982.
6. Scherly, D., Boelens, W., van Venrooij, W.J., Dathan, N.A., Hamm, J. and Mattaj, I.W. (1989) *EMBO J.*, **8**, 4163–4170.
7. Bentley, R.C. and Keene, J.D. (1991) *Mol. Cell. Biol.*, **11**, 1829–1839.
8. Stefano, J.E. (1984) *Cell*, **36**, 145–154.
9. Scherly, D., Boelens, W., Dathan, N.A., van Venrooij, W.J. and Mattaj, I.W. (1990) *Nature*, **345**, 502–506.
10. Hoffman, D.W., Query, C.C., Golden, B L , White, S.W. and Keene, J.D. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 2495–2499.
11. Nagai, K., Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990) *Nature*, **348**, 515–520.
12. Kenan, D.J., Query, C.C. and Keene, J.D. (1991) *Trends Biochem. Sci*, **16**, 214–220.
13. Query, C.C., Deutscher, S.L. and Keene, J.D. Critical role for aromatic residues in the RNA-binding domains of the U1snRNP-70K and the 60kD Ro proteins, submitted, 1991.
14. Wilusz, J. and Keene, J.D. (1986) *J. Biol. Chem.*, 261, 5467–5472.
15. Blackwell, T.K. and Weintraub, H. (1990) *Science*, **250**, 1104–1108.
16. Tuerk, C. and Gold, L. (1990) *Science*, **249**, 505–249.
17. Lutz-Freyermuth, C., Query, C.C. and Keene, J.D. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6393–6397.
18. Scherly, D., Dathan, N.A., Boelens, W., van Venrooij, W.J. and Mattaj, I.W. (1990) *EMBO J.*, **9**, 3675–3681.
19. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) *Science*, **239**, 487–494.
20. Rosenberg, A.H., Lade, B.N., Chui, D.S., Dunn, J.J. and Studier, F.W. (1987) *Gene*, **56**, 125–135.
21. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor University Press, Cold Spring Harbor, pp. 10.32–10.33.
22. Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucleic Acids Res.*, **12(1)**, 387–395.
23. Hope, I.A. and Struhl, K. (1985) *Cell*, **43**, 177–188.
24. Busch, H., Reddy, R., Rothblum, L. and Choi, Y. (1982) *Ann. Rev. Biochem.*, **51**, 617–653.
25. Tindall, K.R. and Kunkel, T.A. (1988) *Biochemistry*, **27**, 6008–6013.
26. Fresco, L.D., Harper, D.S. and Keene, J.D. (1991) *Mol. Cell. Biol.*, **11**, 1578–1589.
27. Lund, E., Kahan, B. and Dahlberg, J.E. (1985) *Science*, 229, 1271–1274.
28. Bach, M., Krol, A. and Luhrman, R. (1990) *Nucleic Acids Res.*, 18, 449–457.
29. Guthrie, C. and Patterson, B. (1991) *Annu. Rev. Genet.*, **22**, 387–419.