The EMBO Journal vol.13 no.16 pp.3873–3881, 1994

NMR studies of U1 snRNA recognition by the N-terminal RNP domain of the human U1A protein

Peter W.A.Howe, Kiyoshi Nagai, David Neuhaus and Gabriele Varani[1]

MRC Laboratory of Molecular Biology, Hills Road, Cambridge
CB2 2QH, UK
[1]Corresponding author

Communicated by I.W.Mattaj

The RNP domain is a very common motif found in hundreds of proteins, including many protein components of the RNA processing machinery. The 70–90 amino acid domain contains two highly conserved stretches of 6–8 amino acids (RNP-1 and RNP-2) in the central strands of a four-stranded antiparallel β-sheet, packed against two α-helices by a conserved hydrophobic core. Using multidimensional heteronuclear NMR, we have mapped intermolecular contacts between the human U1A protein 102 amino acid N-terminal RNP domain and a 31-mer oligonucleotide derived from stem–loop II of U1 snRNA. Chemical shift changes induced on the protein by the RNA define the surface of the β-sheet as the recognition interface. The reverse face of the protein, with the two α-helices, remains exposed to the solvent in the presence of the RNA, and is potentially available for protein–protein contacts in spliceosome assembly or splice site selection. Protein–RNA contacts occur at the single-stranded apical loop of the hairpin, but also in the major groove of the helical stem at neighbouring U·G and U·U non-Watson–Crick base pairs. Examination of a proposed model for the complex in the light of the present results reveals several features of RNA recognition by RNP proteins. The quality of the spectra for this complex of 22 kDa demonstrates the feasibility of NMR investigation of RNA–protein complexes.

Key words: heteronuclear NMR/RNA–protein recognition/structural studies/U1A/U1 snRNA

Introduction

U1 snRNP is one of five RNA–protein particles crucial for pre-mRNA splicing. The human U1 snRNP comprises a 164 nt RNA (U1 snRNA), eight core proteins common to all snRNPs and three U1-specific proteins, U1A, U1C and U1 70K. Base pairing between the 5′-end of U1 snRNA and the 5′-splice site occurs early in spliceosome assembly and is necessary for recognition of that site (Rosbash and Séraphin, 1991; Kohtz *et al.*, 1994). The role of the U1-specific protein factors in splicing is not completely clear. Both U1 70K and U1A appear to be involved in splice-site recognition and selection (Flickinger and Salz, 1994; Kohtz *et al.*, 1994), and U1A may also be involved in linking the splicing and polyadenylation machineries. In addition to its role in splicing, U1A may positively regulate polyadenylation efficiency by interacting with the upstream efficiency element of the SV40 late polyadenylation signal (Lutz and Alwine, 1994). Expression of U1A is regulated by negative feedback; excess U1A binds an RNA element within the 3′-untranslated region (UTR) of its own mRNA and inhibits polyadenylation by interacting with the poly(A)-polymerase (Boelens *et al.*, 1993; van Gelder *et al.*, 1993; Gunderson *et al.*, 1994). The U1A protein has very high affinity for U1 snRNA stem–loop II ($K_d \approx 10^{-11}$ M) (Hall and Stump, 1992; van Gelder *et al.*, 1993), and binds U1 snRNA immediately after the RNA has been transcribed (Terns *et al.*, 1993). By contrast, U1 70K and U1C only bind the RNA when it returns to the nucleus after processing in the cytoplasm.

Both U1A and U1 70K contain at least one RNP domain [also called RNA-binding domain (RBD) or RNA recognition motif (RRM)], a motif found in hundreds of RNA-binding proteins including snRNP components, splicing factors and regulators, hnRNP proteins and factors involved in mRNA 3′-end formation (Mattaj, 1993). The 283 amino acid human U1A protein comprises two RNP domains. The N-terminal domain (amino acids 4–98) recognizes the highly conserved loop II of U1 snRNA with the same high affinity as does the entire U1A protein (Scherly *et al.*, 1990; Jessen *et al.*, 1991), although the basic region between amino acids 100 and 115 may further modulate the specificity (Scherly *et al.*, 1991). The second RNP domain (amino acids 210–280) and the region connecting the two RNP domains are not required for U1 snRNA recognition, but are necessary for nuclear localization (Kambach and Mattaj, 1992) and regulation of the polyadenylation of the U1A pre-mRNA (Boelens *et al.*, 1993; Gunderson *et al.*, 1994) . The modular structure of U1A is shared by many RNA-binding proteins of the RNA processing machinery (Biamonti and Riva, 1994).

The crystal structure of the A95 fragment (residues 2–95) of the human U1A protein (Nagai *et al.*, 1990), together with NMR studies of a similar fragment (Hoffman *et al.*, 1991) and of hnRNP C (Wittekind *et al.*, 1992), revealed the characteristic fold of the RNP domain. The highly conserved RNP-1 and RNP-2 sequences that define the domain are found in the two central strands of a four-stranded antiparallel β-sheet packed through a hydrophobic core against two α-helices. However, the structures of isolated RNP-containing proteins have revealed little about the mechanism of RNA recognition, because the same RNP structure can recognize very different RNA structures. In particular, the N-terminal domain from U1A recognizes with similar affinity stem–loop II of U1 snRNA (Scherly *et al.*, 1990) and a bulge region within the 3′-UTR of its pre-mRNA (van Gelder *et al.*, 1993), hnRNP C binds

© Oxford University Press

poly(U) (Wittekind *et al.*, 1992), and hnRNP A1 binds a single-stranded, purine-rich sequence (Burd and Dreyfuss, 1994). More extensive structural studies on complexes between proteins containing the RNP domain and their cognate RNAs are necessary for detailed understanding of the mechanism of recognition.

In this report, we describe the NMR investigation of the complex between a 31-mer oligonucleotide derived from U1 snRNA stem–loop II, and the 102 amino acid N-terminal fragment of the human U1A protein. In the case of the complex between hnRNP C and poly(U), NMR studies have demonstrated that hnRNP C recognizes its cognate rU_8 RNA via amino acids located on the surface of the β-sheet (Görlach *et al.*, 1992). As with hnRNP C (Görlach *et al.*, 1992), and as proposed from site-directed mutagenesis and model building (Jessen *et al.*, 1991), we show here that the surface of the antiparallel β-sheet is the recognition interface between U1A and stem–loop II of U1 snRNA. The sites of protein–RNA contacts are found both in the hairpin loop, consistent with extensive biochemical data, and in the major groove of the helical stem at the structural distortion induced by consecutive non-Watson–Crick base pairs.

Results

Protein–RNA contacts occur in the RNA major groove

The sequence and secondary structure of the 31-mer oligonucleotide from U1 snRNA stem–loop II is shown in the insert at the top of Figure 1A. Initial experiments were conducted on a 27-mer wild-type fragment from U1 snRNA. However, UV melting and NMR experiments showed that the RNA forms predominantly a dimerized duplex structure instead of the desired hairpin at the millimolar concentrations required for NMR. Dimerization occurs because a very stable duplex is formed between the loop region (nucleotides $G_{69}-C_{76}$) and a complementary sequence within the stem (nucleotides $G_{77}-C_{83}$) of the wild-type loop II sequence. Twelve consecutive Watson–Crick base pairs, interrupted by a single bulged U residue, are present in this very stable dimer structure. To overcome this problem, two successive Watson–Crick base pairs were reversed ($U_{63}{\cdot}A_{78}$ to $A_{63}{\cdot}U_{78}$ and $C_{64}{\cdot}G_{77}$ to $G_{64}{\cdot}C_{77}$) and four unpaired adenines were added at the 3′-end of the hairpin to improve transcriptional efficiency. Reversal of the $U_{63}{\cdot}A_{78}$ and $C_{64}{\cdot}G_{77}$ base pairs destabilizes the duplex without significantly affecting the stability of the hairpin form of the RNA. As expected, the mutant sequence forms the desired hairpin structure, as demonstrated both by NMR and by the concentration dependence of the UV melting profiles (data not shown). As demonstrated by two-dimensional NMR, the four unpaired adenines at the 3′-end of the hairpin do not perturb the structure of the stem–loop. In biochemical studies of the cross-reactivity of U1A and U2B″ proteins with their cognate RNAs, the stem sequences contributed little to binding affinity or specificity, although a Watson–Crick paired stem was required for high-affinity binding (Scherly *et al.*, 1990). Consistent with these results, the wild-type and quadruple-mutant RNA sequences bind the A102 protein (amino acids 2–102 of the human U1A protein) with comparable affinity.

NMR spectra of the downfield region of the 31-mer RNA oligonucleotide derived from the human U1 snRNA loop II sequence, and its complex with the A102 protein domain, are shown in Figure 1A. Analysis of the NMR spectra confirmed the base pairing expected for the stem region, and revealed the formation of two consecutive non-Watson–Crick pairs ($U_{61}{\cdot}G_{80}$ and $U_{60}{\cdot}U_{81}$). Assignment of the exchangeable, aromatic and H1′ resonances for the free RNA from the base-paired stem and the well structured

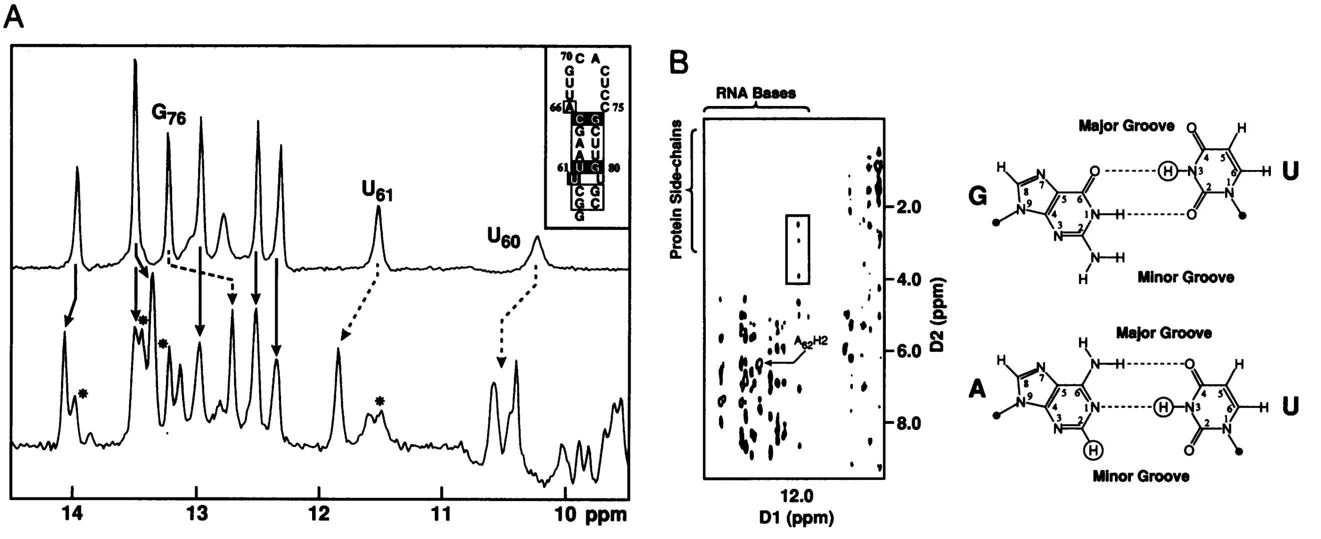

Fig. 1. (**A**) One-dimensional NMR spectra at 10°C and in 5 mM phosphate buffer of loop II RNA (top) and of its complex with the A102 protein (bottom). In the sample corresponding to the spectrum shown at the bottom, the RNA is present at ~30% excess at a total concentration of ~1.2 mM. The stoichiometry is confirmed by the relative intensities of the peaks corresponding to the free (starred resonances) and bound forms of the RNA in the spectrum shown at the bottom. Nucleotides whose exchangeable NH or A H2 resonances are shifted significantly (>0.2 p.p.m.) by the protein are shown black in the insert, whereas those that are essentially unaffected (chemical shift changes <0.2 p.p.m.) are shown boxed. (**B**) Two-dimensional NOESY spectrum of the same complex as in panel A. The characteristically shifted A_{62} H2 resonance is identified by its strong cross-peak to U_{79} NH. The intermolecular NOE contacts between U_{61} NH and amino acid side chains are boxed. Drawings of the GU and AU base pairs show the NMR reporter nuclei in the major and minor grooves.

portion of the loop ($A_{66}-U_{68}$), was accomplished using unlabelled and ^{15}N-labelled RNA and two-dimensional NMR methods (Varani and Tinoco, 1991). Two relatively sharp imino resonances are observed at 10.3 and 11.6 p.p.m.; two broader resonances are also observed at lower pH and temperature. Both the ^{1}H and ^{15}N chemical shifts of these four upfield-shifted imino resonances are consistent with what was observed for U·G and U·U base pairs. Although the strong intra-base-pair NOEs expected for the $U_{61}{\cdot}G_{80}$ and $U_{60}{\cdot}U_{81}$ base pairs are not present, this is probably because the imino protons exchange too rapidly with the solvent. This is often the case for non-Watson–Crick base pairs in RNA, and in this particular case the two consecutive mismatches are expected to enhance helix opening and increase the rate of exchange. Other NOE interactions, and the chemical shifts of the imino protons, support the presence of these base pairs, in agreement with the results of chemical and enzymatic mapping of the full-length human U1 snRNA (Krol *et al.*, 1990). The A_{62} H2 resonance, adjacent to the $U_{61}{\cdot}G_{80}$ pair, resonates at 6.3 p.p.m., ~1 p.p.m. upfield from what is generally observed for a perfect helix, suggesting an unusual conformation induced by the non-Watson–Crick pairs. The observation of characteristic NOE cross-peaks shows that base stacking is continued from the stem to A_{66}-U_{67}-U_{68} in the nominally single-stranded loop. The first residue in the loop, A66, is base-paired to a uracil residue, presumably U_{73}. Other loop resonances are broadened by conformational exchange, suggesting that the remainder of the loop is flexible, although partially stacked, in the absence of the protein.

Upon addition of the protein to form a sample where RNA is present in 30% excess, two separate sets of resonances are observed (Figure 1A, bottom spectrum). These correspond to free RNA resonances (identified by stars in the bottom spectrum) and bound RNA resonances. Thus, the complex is in slow exchange on the NMR time scale. If the on rate for complex formation is diffusion limited, slow exchange on the NMR time scale implies a half-life for the complex of >1 ms ($k_{\mathrm{off}} >> 10^{-3}$ s), as expected from the sub-nanomolar dissociation constant ($K_{\mathrm{d}} \approx 10^{-11}$ M) (Hall and Stump, 1992; van Gelder *et al.*, 1993).

In the presence of the protein, most imino and H2 resonances shift by 0.1 p.p.m. or less, but G_{76} NH shifts 0.7 p.p.m. upfield and U_{61} NH shifts 0.5 p.p.m. downfield (Figure 1A). This result reveals that the protein contacts not only the apical loop, but also the region of the stem neighbouring the consecutive non-Watson–Crick base pairs. A similar conclusion was reached from the pattern of phosphate ethylation protection (Jessen *et al.*, 1991). Several new resonances appear in the complex at the upfield edge of the RNA imino proton region, between 9.5 and 10.5 p.p.m. These resonances are due both to non-base-paired RNA imino resonances that are protected from exchange with the solvent only in the presence of the bound protein, and to protein NH resonances downfield-shifted upon RNA binding.

Crystal structures of tRNA-synthetase complexes show protein–RNA contacts occurring in the wide and shallow minor groove of RNA helices (Rould *et al.*, 1989). Uracil NH and adenine H2 resonances act as ideal reporter signals in the minor groove (Figure 1B). If the N-terminal RNP domain of the U1A protein were contacting the RNA bases in the minor groove, one would expect significant chemical shift changes for A H2 and U NH resonances from the two consecutive A·U pairs, and the observation of intermolecular RNA–protein NOE cross-peaks. Since none of the A H2 or U NH resonances, including A_{66} at the stem–loop junction, are significantly shifted upon addition of the protein (Figure 1B), RNA recognition is unlikely to occur in the minor groove. Further, the sharp and unusually shifted A_{62} H2 resonance at 6.3 p.p.m. (Figure 1B) would be a very convenient resonance to identify intermolecular NOE contacts involving the RNA minor groove. We have been unable to find any contact between protein side chains and this well resolved resonance, or any of the other A H2 resonances in the more crowded spectral region between 7 and 8 p.p.m.

Although the major groove of regular A-form RNA is too narrow to allow insertion of protein secondary structure elements (Steitz, 1990), non-Watson–Crick base pairs may widen the major groove to allow recognition of functional groups on the bases (Weeks and Crothers, 1991). Several well resolved NOE cross-peaks occur between protein side chains and U_{61} NH (identified from the chemical shift of its attached nitrogen) in the major groove of the wobble $U_{61}{\cdot}G_{80}$ pair (Figure 1B, boxed cross-peaks). The intermolecular NOE interactions between the protein side chains and RNA resonances in the major groove, together with the pattern of chemical shift changes, clearly identify the face of the stem–loop defined by the major groove and its extension into the hairpin loop as the recognition interface.

RNA recognition occurs on the surface of the β-sheet and at a newly identified helix after the RNP domain

Nearly complete spectral assignments for the backbone ^{1}H and ^{15}N resonances of the A102 protein domain have been obtained using unlabelled and ^{15}N-labelled protein. Spectral assignments were obtained at 42°C and 20°C using well established procedures (Wüthrich, 1986), and will be reported elsewhere. The present assignments are generally consistent with those previously reported for the A95 fragment (Hoffman *et al.*, 1991). With the exception of a few prolines, the only gaps in the assignments occur at amino acids 47–49 and 88–92 (Figure 2): these regions are flexible in the crystal (Nagai *et al.*, 1990) and appear to be flexible in solution as well. Analysis of the NOE interactions (Figure 2) and chemical shifts from the A102 protein domain confirms the secondary structure previously observed by crystallography (Nagai *et al.*, 1990) and NMR (Hoffman *et al.*, 1991). In addition, the i, i+3 NOE connectivities and the close contacts between amide protons on consecutive amino acids that are diagnostic of helical regions in proteins, reveal the presence of a third helix, extending at least from amino acid 93 to 97. The chemical shift index (CSI) (Wishart *et al.*, 1992) can be used to provide a qualitative indication of the α-helical or β-sheet character of a particular amino acid. In the case of the A102 protein, the results of the CSI analysis are satisfactorily consistent with the secondary structure revealed by X-ray crystallography and confirmed by the pattern and intensities of the NOE cross-peaks.

The presence of an additional helix just after the C-

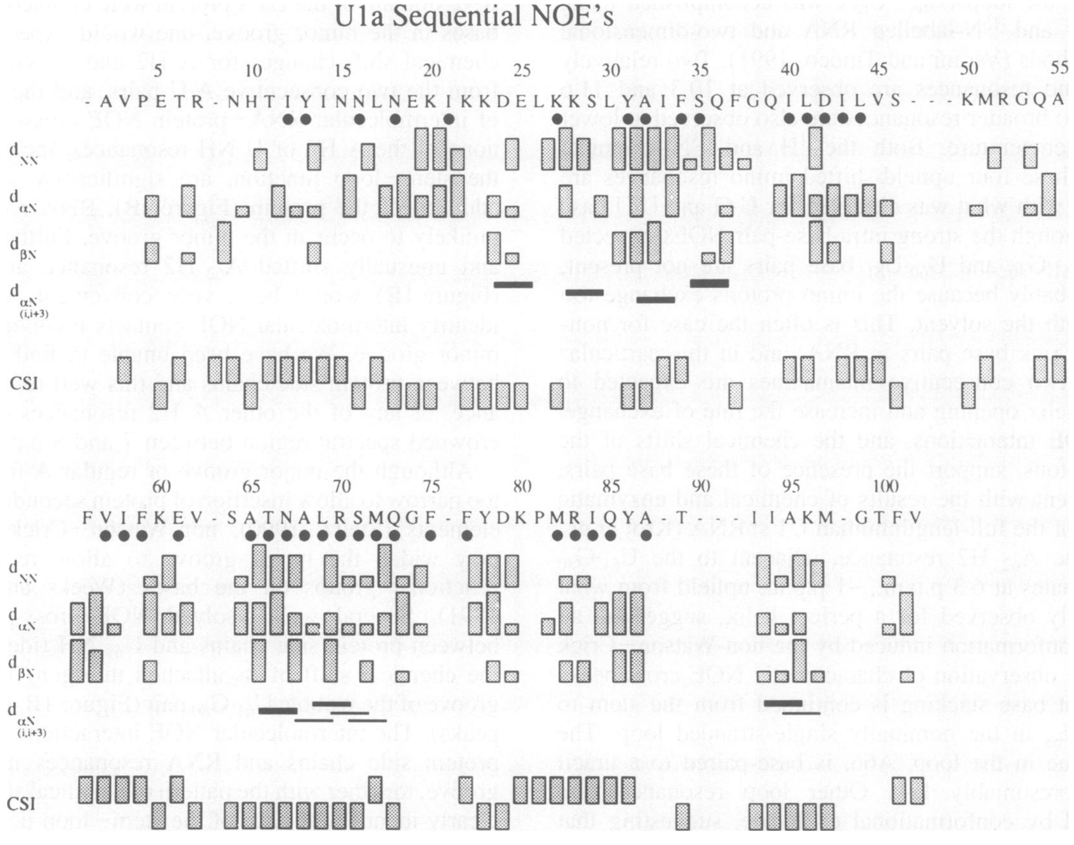

Fig. 2. Sequential NOE signals for assigned backbone resonances of the A102 protein domain (amino acids 2–102 of the human U1A protein). Gaps in the sequence correspond to those few amino acids, generally from two flexible loop regions, for which reliable assignments are still unavailable. The intensities of the sequential NOE cross-peaks are indicated schematically by the height of the bars, and slowly exchanging amide resonances by filled circles. The chemical shift index (CSI) (Wishart *et al.*, 1992) is also reported.

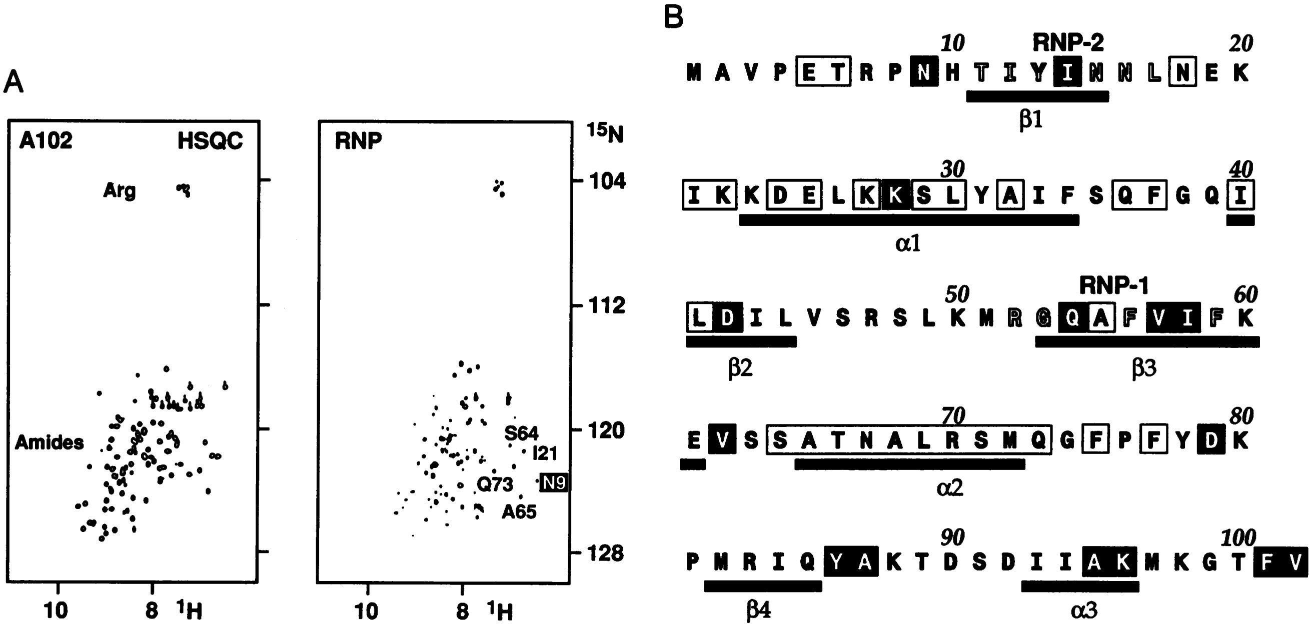

Fig. 3. (A) Two-dimensional ^{1}H$-^{15}$N (HSQC) correlated spectra recorded at 20°C for the free A102 protein (left) and the complex with its cognate stem–loop II RNA (right). A few representative resonances that are either shifted (N$_9$) or not shifted (all others) by the RNA are highlighted. (B) Protein sequence and secondary structure. Resonances from amino acids drawn as white on black letters are significantly shifted by the RNA (>0.2 p.p.m. in the ^{1}H dimension or >1 p.p.m. in the ^{15}N dimension), whereas resonances from boxed amino acids are shifted by smaller amounts or unchanged.

terminus of the RNP domain could not have been observed in previous studies because the U1A fragments studied in the past were truncated at residue 95. Since amino acids 95–98 are necessary for RNA binding (Jessen *et al.*, 1991; Scherly *et al.*, 1991), this third helix is likely to be important for RNA recognition. Consistent with this proposal, the intensity of NOE cross-peaks from the C-terminal residues increases in the presence of the RNA, suggesting the stabilization of the local secondary structure upon binding to stem–loop II RNA.

Since we have obtained nearly complete assignments for NH resonances of the free protein, the protein resonances in contact with the RNA can be identified by comparing $^{1}H-^{15}N$ correlation spectra of the free and bound A102 protein (Figure 3A). Assignments for the bound form of the protein were obtained by comparing HSQC (Figure 3A), HSQC-NOESY (Gronenborn *et al.*, 1989) and ^{15}N-half-filtered NOESY (Otting and Wüthrich, 1990) spectra for the free and bound ^{15}N-labelled A102 protein. The cut-off between significantly and non-significantly shifted resonances was set at 0.2 p.p.m. for ^{1}H and 1 p.p.m. for ^{15}N. Chemical shift changes of similar magnitudes have been observed at the DNA interface of the homeodomain–DNA complex (Qian *et al.*, 1993). Chemical shifts for the free and bound form of the protein could be compared for ~50% of all protein backbone NH resonances, as well as several side chain amide resonances. Spectral overlap, limited sensitivity and the changes induced by the RNA prevented reliable assignments for the remainder of the protein in the complex, and ^{13}C-labelling will be required to complete the spectral assignments for the complex.

With the exception of K_{28}, all unambiguously assigned resonances from the first two α-helices are essentially unaffected by the RNA, while several resonances from the C-terminal region and the β-sheet are significantly shifted (Figure 3B). Interpreting these results in terms of the crystal structure of the free A95 domain, clearly defines the surface of the β-sheet as the recognition interface (Figure 4A). The opposite face of the protein, where the first two α-helices are located, remains exposed to solvent when the RNA is bound. The very small perturbation of resonances in these two α-helices, and the conservation of some characteristic long-range NOE interactions, demonstrate that the structure of the RNP domain is largely unaltered in the complex. Minor structural changes could explain the few changes in chemical shifts in loops connecting secondary structure elements that reside away from the main surface of intermolecular contacts. Resonances within the proposed third helix (A_{95} and K_{96}) and at the end of the protein fragment ($T_{100}-V_{102}$) are also affected by the RNA, confirming that the newly identified helix is involved in RNA recognition.

Examination of the model for U1A–loop II recognition

The present results confirm and extend the main features of the existing model for the U1A–loop II complex (Jessen *et al.*, 1991). The fold of the RNP domain is unaltered in the presence of the RNA, and the chemical shift changes induced by the RNA binding on the protein provide direct physical evidence that the surface of the β-sheet is the site of intermolecular contacts. Almost all the protein backbone and side chain NH signals that are affected by the RNA are at the RNA–protein interface in the existing model (Figure 4B). In contrast, most of the unaffected resonances are on the opposite face of the protein (Figure 4A). The only amino acid from the face of the protein opposite the RNA whose resonances are significantly shifted in the complex is K_{28}. A similarly located residue in hnRNP C, K_{30}, was similarly affected by rU_8 binding (Görlach *et al.*, 1992), suggesting a general but at present unknown role for this lysine residue in RNA recognition by RNP domains.

The results of site-directed mutagenesis complement the NMR results. Amino acid mutations which reduce RNA binding generally map to side chains not yet assigned in the complex, but located near amino acids whose assigned signals are affected by the RNA (Figure 4B). Most side chain substitutions that reduce binding significantly, for instance T_{11}, N_{15}, R_{52} or $Q_{54,}$ are on the surface of the β-sheet. The present data confirm that loss in binding results from the loss or perturbation of intermolecular contacts, rather than protein misfolding.

The intermolecular NOE interactions involving the $U_{61}{\cdot}G_{80}$ base pair, and the pattern of chemical shift changes, demonstrate that the face of the RNA contacted by the protein is the major-groove side of the short helix defined by the RNA stem. In the free RNA, sequential NOE connectivities show that A-form-like base stacking is continued from the stem into the single-stranded loop to A_{66}-U_{67}-U_{68}; after this residue, the sequential NOE connectivities are clearly interrupted. Although the broadening of the base resonances suggest the presence of base stacking interactions in the loop, the $G_{69}-C_{76}$ single-stranded sequence is in conformational exchange in the absence of the protein. A large conformational change in the RNA structure upon protein binding can be ruled out by the very small chemical shift changes observed for most stem resonances upon complex formation, but protein binding may well stabilize a preferred conformation within the envelope of structures sampled by the free RNA. The model proposes that the direction of the phosphodiester backbone is reversed after U_{68} to bridge the two sides of the stem while maximizing base stacking, as observed in the crystal structures of tRNA anticodon loops (Krol *et al.*, 1990). In this case, all phosphates protected by U1A from ethylation lie on the major-groove side of the RNA helix. The present results demonstrate that it is the major groove and its extension into the hairpin loop which is the face of the RNA recognized by the U1A protein.

Discussion

Despite the importance of RNA–protein complexes in many cellular functions, very little structural information is available on the mechanism of RNA recognition (Mattaj, 1993). The only notable exceptions are the crystal structures of tRNAs complexed with their cognate synthetases (Rould *et al.*, 1989; Caverelli *et al.*, 1993; Biou *et al.*, 1994), and the NMR study of the complex between hnRNP

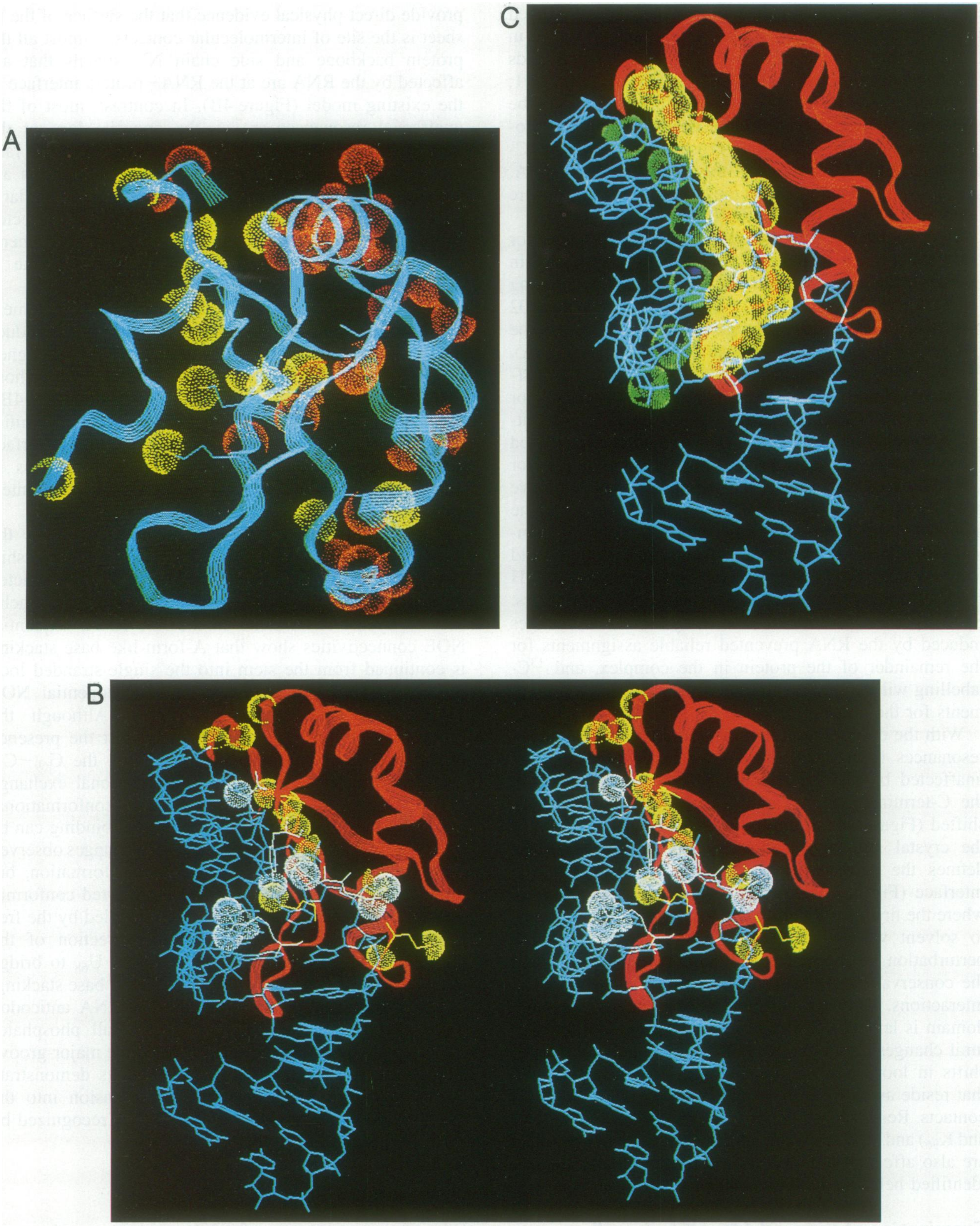

Fig. 4. (**A**) Protein resonances that are shifted significantly upon complex formation arise from residues (shown in yellow) that are generally located on the surface of the β-sheet in the crystal structure (Nagai *et al.*, 1990) of the A95 fragment. In contrast, unaffected resonances arise from residues (in red) that are generally located on the opposite face of the protein. (**B**) Stereo view of the model of the complex between the A95 protein domain and stem–loop II. The RNA hairpin is shown from the major groove side. Amino acids whose amide NMR signals are shifted by the RNA are shown in yellow, whereas white van der Waals surfaces identify the location of side chains defined by site-directed mutagenesis as important for RNA recognition (Nagai *et al.*, 1990; Scherly *et al.*, 1990; Jessen *et al.*, 1991). (**C**) The structural complementarity between the stacked RNA backbone (phosphates are identified by green van der Waals surfaces) and the protein antiparallel β-sheet (yellow) is highlighted in this view of the model.

C and rU_8 (Görlach *et al.*, 1992). In the present study, we have followed the same NMR approach used for the hnRNP C–rU_8 complex by Görlach and co-workers (Görlach *et al.*, 1992) to identify the recognition interface between the human U1A protein N-terminal RNP domain and its cognate RNA, stem–loop II of human U1 snRNA.

Several novel features of RNA recognition by RNP domains emerge from the examination of the present results, and some may have more general implications for RNA recognition by RNA-binding proteins. Hairpin loops, internal loops and bulges are common targets for recognition by RNA-binding proteins. In this respect, the complexes between the U1A N-terminal RNP domain and its cognate RNAs, stem–loop II from U1 snRNA and the internal loop within the 3′-UTR of the U1A pre-mRNA, provide ideal paradigms for RNA–protein interactions.

The RNP domain is able to recognize very different RNA structures with high affinity and specificity. In the hnRNP C–rU_8 interaction (Görlach *et al.*, 1992), the recognition interface between the RNP domain and the cognate RNA is the surface of the β-sheet. Similarly, recognition of stem–loop II by the U1A protein occurs via amino acids located on the surface of the β-sheet, and in a newly identified α-helix located at the C-terminus of the domain (Figure 4A). The similar right-handed twist and characteristic repeat of antiparallel β-sheets and of the nucleic acid phosphodiester backbone (Carter and Kraut, 1974) suggest an intrinsic complementarity that may be an important factor in RNA recognition by RNP and other RNA-binding proteins. The structural complementarity between the RNA backbone of the loop II RNA and the A102 β-sheet surface (Figure 4C) may favour electrostatic contacts and hydrogen bond interactions between protein side chains and the RNA phosphates. Amino acids exposed on the surface of the β-sheet could provide specific contacts with the functional groups on the bases that are required for sequence discrimination.

Unlike DNA–protein recognition, where very large changes in affinity are observed upon single base substitutions, there is no known point mutation in the loop II RNA which leads to a loss in binding greater than ~10-fold (Jessen *et al.*, 1991; Hall and Stump, 1992). Rather than sequence alone, it is likely that RNA structure may be primarily responsible for sequence-specific recognition. In this respect, it is interesting to observe that the same stacking pattern is observed in the single-stranded region of stem–loop II within U1 snRNA and in the high affinity site for the U1A N-terminal domain in the 3′-UTR of the U1A pre-mRNA (van Gelder *et al.*, 1993), despite the very different secondary structural context. Furthermore, the surface of the β-sheet appears again to represent the recognition interface (C.C.Gubser and G.Varani, unpublished results). The structural complementarity between the RNA backbone and the antiparallel β-sheet, rather than direct contacts between the protein side-chains and the RNA bases, may thus provide a large component of the free energy of recognition between U1A and loop II. This complementarity, which is the direct result of the stereochemistry of the protein and RNA, may be the structural basis for the ability of RNP domains to recognize such a remarkable range of RNA structures.

Unlike previous studies on peptide models of other RNA–protein complexes, the entire surface of the β-sheet in the folded protein is involved in RNA recognition, including the highly conserved 6–8 amino acid RNP-consensus repeats, and the α-helical region at the C-terminus of the polypeptide (Figure 4B). As observed for tRNA-synthetase complexes (Rould *et al.*, 1989; Caverelli *et al.*, 1993; Biou *et al.*, 1994), it is not possible to single out a short polypeptide from the A102 protein as the single site of intermolecular recognition.

The loop connecting β2 and β3 is one of the most variable regions among different RNP domains, and it is a very important determinant of RNA discrimination for U1A and U2B″ (Scherly *et al.*, 1990). The β2–β3 loop is variable both in sequence and size between different RNP domains. In the model of the U1A—stem–loop II complex, that loop can be inserted in the RNA major groove and be extended to easily reach the non-Watson–Crick $U_{61}{\cdot}G_{80}$ base pair. At the same time, residues located within β2 and β3 can be positioned in the vicinity of the three or four terminal nucleotides of the single-stranded loop (Jessen *et al.*, 1991), where the sequences of the RNA cognates of U1A and U2B″ diverge. Unfortunately, the loop is very flexible even in the crystal of the isolated protein, and its high flexibility in solution has so far prevented reliable assignments and structural characterization, and the consequent identification of intermolecular contacts.

The protein secondary structure C-terminal to the RNP domain has not been reported before. The highly basic region of the protein following the RNP domain is important for U1 snRNA recognition by U1A (Scherly *et al.*, 1991) and poly(U) recognition by hnRNP C (Burd and Dreyfuss, 1994). The sequence homology between human and frog U1A proteins has led to the suggestion that this region is an integral part of the RNA recognition domain (Scherly *et al.*, 1991) for at least some members of the RNP family of proteins. The structure of the A117 polypeptide (amino acids 2–117 of the human U1A protein) is currently being investigated to determine the location of this α-helix more precisely, and to identify its structural relationship with respect to the core of the RNP domain.

In contrast with the study of the complex between hnRNP C and poly(U), stem–loop II is a well-defined hairpin structural element. Therefore, RNA structure, rather than sequence alone, can be expected to be important for sequence-specific recognition. As also observed in the anticodon loops of tRNAs and many RNA secondary structure elements (Varani and Tinoco, 1991), helical stacking is continued from the base-paired stem of stem–loop II into the single-stranded loop. Although, not surprisingly, the loop is quite flexible and in conformational exchange for the isolated RNA, protein binding may stabilize the loop conformation. Folding at the nucleic acid interface of flexible protein domains often contributes a significant fraction of the overall free energy of protein–DNA recognition (Spolar and Record, 1994). In this complex, our data suggest that most of the protein interface is well-ordered in the absence of the RNA, but protein-induced folding of the flexible RNA loop into a unique and stable structure may be important for U1A–loop II interaction.

The deep A-form RNA major groove is too narrow to allow base recognition by insertion of protein secondary structure elements (Steitz, 1990), but the consecutive non-Watson–Crick base pairs ($U_{61}\cdot G_{80}$ and $U_{60}\cdot U_{81}$) may open up the major groove to allow recognition, as originally proposed for the HIV Tat–TAR interaction (Weeks and Crothers, 1991). The contacts between the protein side chains and the major groove of the RNA in the vicinity of the $U_{60}\cdot U_{81}$ and $U_{61}\cdot G_{80}$ base pairs may provide a structural explanation for the reduced affinity (~10-fold) of U1A for an RNA containing a perfectly paired stem (Jessen *et al.*, 1991).

Most DNA-binding proteins contact DNA bases by inserting secondary structure elements, most often α-helices, into the wide B-form DNA major groove (Steitz, 1990), but the TATA-box binding protein provides a notable counter-example, since it recognizes the TATA-box minor groove (J.L.Kim *et al.*, 1993; Y.Kim *et al.*, 1993). Although the biochemical evidence in support of RNA recognition in the major groove is well established for several RNA–protein complexes (Weeks and Crothers, 1991; Hamy *et al.*, 1993), the present results provide direct physical evidence in support of this potentially general aspect of RNA recognition. Thus, there are now examples of DNA- and RNA-binding proteins that recognize their cognates in either the major or the minor grooves.

Conclusions

The present NMR study of the human U1A N-terminal RNP domain complexed with a U1 snRNA loop II oligonucleotide provides direct physical evidence in support of the existing model for the RNA–protein complex (Jessen *et al.*, 1991). Intermolecular contacts occur on the surface of the antiparallel β-sheet of the RNP domain, whereas the opposite face of the protein is exposed to the solvent and therefore potentially available for protein–protein contacts in spliceosome assembly or splicing regulation. The RNA is contacted in the major groove of the short helix in the vicinity of consecutive non-Watson–Crick base pairs, and on the face of the hairpin loop corresponding to the major groove. RNA recognition in the major groove at or near structural distortions induced by internal loops, bulges or non-Watson–Crick base pairs may be a very general mechanism of RNA recognition by RNA-binding proteins (Weeks and Crothers, 1991). The antiparallel β-sheet and the stacked phosphodiester backbone define a general complementary interface for RNA recognition by RNP domains. The sequence-specific contacts with the RNA bases that must take place to define the sequence specificity of different RNP domains will be identified for this complex when a high-resolution structure is determined by crystallography and NMR.

High quality NMR spectra have been obtained for this complex between the folded protein domain and a non double-helical nucleic acid structure. With the exception of the investigation of the hnRNP C–rU_8 complex (Görlach *et al.*, 1992), NMR studies of RNA–protein recognition have hitherto been limited to the individual RNA (Colvin *et al.*, 1993; Hoffman *et al.*, 1993; Jaeger and Tinoco, 1993; Wimberly *et al.*, 1993) or protein (Hoffman *et al.*, 1991; Wittekind *et al.*, 1992) components, or to small peptides in complex with oligonucleotides (Puglisi *et al.*, 1992, 1993). This study demonstrates that NMR can be used to investigate the recognition of structured RNAs by folded protein domains, without the need to reduce either the protein or the RNA to its minimal recognition elements. The molecular weight of 22 kDa, and the difficulties of NMR investigation of RNA single-stranded regions (Varani and Tinoco, 1991), combine to make the determination of this structure a very challenging task. Isotopic labelling of both the RNA and protein components will be required to obtain a high resolution structure of the complex, and such studies are already under way. The remarkable quality of the spectra, and the ability to label isotopically either the RNA or protein components, promise a structure of very high quality.

Materials and methods

Protein purification

The A102 fragment (residues 2–102 of the human U1A protein) was expressed in *Escherichia coli* using the T7 RNA polymerase expression system and purified using CM-Sepharose as previously described (Nagai *et al.*, 1990). ^{15}N-labelling the protein to a level of >98% was achieved by growing *E.coli* in the presence of $^{15}NH_4Cl$ as the sole nitrogen source. Mass spectrometry confirmed the purity of the sample and revealed that the N-terminal methionine is cleaved in this construct.

RNA preparation and characterization

The oligonucleotide 5′-pppGGCUUA**A**GCAUUGCACUCCGG**U**UGU-GCA**AAA** (bold characters denote differences with the human U1 snRNA loop II sequence) was synthesized using T7 RNA polymerase and synthetic DNA templates (Milligan *et al.*, 1987) containing a double-stranded 17 bp promoter region and a single-stranded template. The RNA was purified by 20% polyacrylamide gel electrophoresis, electroeluted (Schleicher & Schuell), desalted on Sephadex G-15 and extensively dialysed against the NMR buffer (5 mM phosphate buffer, pH 5.5). The sequence was verified enzymatically (Donis-Keller *et al.*, 1977), and RNA concentrations were determined from the UV absorbance at 260 nm (Puglisi and Tinoco, 1990). UV-melting experiments (Puglisi and Tinoco, 1990) were recorded on a Gilson UV-VIS temperature-controlled spectrophotometer by raising the temperature from 2°C to 90°C at a heating rate of 0.5°C per minute. The samples were pre-heated to 90°C and subsequently slowly cooled to 2°C. Cells of different path length were used to obtain melting curves over a 200-fold concentration range (~2 μM to ~400 μM).

NMR spectroscopy

All spectra were recorded using a Bruker AMX-500 NMR spectrometer equipped with a triple resonance probe and operating at 500.14 MHz for ^{1}H and 50.68 MHz for ^{15}N. One-dimensional ^{1}H-NMR spectra were recorded at ~1.2 mM RNA concentration in 5 mM phosphate buffer at pH 6 using the jump-return scheme for water suppression (Plateau and Guéron, 1982). Correlated $^1H-^{15}N$ HSQC spectra (Bodenhausen and Ruben, 1980) were obtained for the free and bound protein (the sample concentration was ~1 mM for the complex) at 20°C in 5 mM D_3-acetate buffer, pH 4.8. Water suppression was achieved using spin-lock purge pulses (Messerle *et al.*, 1989), combined with extremely low power selective saturation during the relaxation delay. ^{15}N-HSQC, HSQC-NOESY (Gronenborn *et al.*, 1989) and ^{15}N-half-filtered NOESY (Otting and Wüthrich, 1990) spectra were recorded for the free and bound ^{15}N-labelled A102 protein. A total of 256 increments was collected, with spectral widths of 8064 Hz (^{1}H) or 5000 Hz (^{15}N), and NOE mixing times of 100 ms (for the RNA–protein complex) or 200 ms (free protein and RNA).

Acknowledgements

It is a pleasure to acknowledge Chris Oubridge, Frederic Allain, Ian Fernley and Charles Gubser for their help at various stages of this project. It is also a pleasure to thank Prof. I.Tinoco,Jr, University of California at Berkeley, in whose laboratory the initial characterization of the RNA by UV melting experiments was carried out. P.H. was supported by an MRC pre-doctoral scholarship.

References

Biamonti,G. and Riva,S. (1994) *FEBS Lett.*, **340**, 1–8.

Biou,V., Yaremchuk,A., Tukalo,M. and Cusack,S. (1994) *Science*, **263**, 1404–1410.

Bodenhausen,G. and Ruben,D.J. (1980) *Chem. Phys. Lett.*, **69**, 185–189.

Boelens,W.C., Jansen,E.J.R., van Venrooij,W.J., Stripecke,R., Mattaj,I.W. and Gunderson,S.I. (1993) *Cell*, **72**, 881–892.

Burd,C.G. and Dreyfuss,G. (1994) *EMBO J.*, **13**, 1197–1204.

Carter,C.W.J. and Kraut,J. (1974) *Proc. Natl Acad. Sci. USA*, **71**, 283–287.

Caverelli,J., Rees,B., Ruff,M., Thierry,J.-C. and Moras,D. (1993) *Nature*, **362**, 181.

Colvin,R.A., White,S.W., Garcia-Blanco,M.A. and Hoffman,D.W. (1993) *Biochemistry*, **32**, 1105–1112.

Donis-Keller,H., Maxam,A.M. and Gilbert,W. (1977) *Nucleic Acids Res.*, **4**, 2527–2538.

Flickinger,T.W. and Salz,H.K. (1994) *Genes Dev.*, **8**, 914–925.

Görlach,M., Wittekind,M., Beckman,R.A., Mueller,L. and Dreyfuss,G. (1992) *EMBO J.*, **11**, 3289–3295.

Gronenborn,A.M., Bax,A., Wingfield,P.T. and Clore,G.M. (1989) *FEBS Lett.*, **243**, 93–98.

Gunderson,S.I., Beyer,K., Martin,G., Keller,W., Boelens,W.C. and Mattaj,I.W. (1994) *Cell*, **76**, 531–541.

Hall,K.B. and Stump,W.T. (1992) *Nucleic Acids Res.*, **20**, 4283–4290.

Hamy,F., Asseline,V., Grasby,J., Iwai,S., Pritchard,C., Slim,G., Butler,P.J.G., Karn,J. and Gait,M. (1993) *J. Mol. Biol.*, **230**, 111–123.

Hoffman,D.W., Query,C.C., Golden,B.L., White,S.W. and Keene,J.D. (1991) *Proc. Natl Acad. Sci. USA*, **83**, 2495–2499.

Hoffman,D.W., Colvin,R.A., Garcia-Blanco,M.A. and White,S.W. (1993) *Biochemistry*, **32**, 1096–1104.

Jaeger,J.A. and Tinoco,I.J. (1993) *Biochemistry*, **32**, 12522–12530.

Jessen,T.H., Oubridge,C., Teo,C.H., Pritchard,C. and Nagai,K. (1991) *EMBO J.*, **10**, 3447–3456.

Kambach,C. and Mattaj,I.W. (1992) *J. Cell Biol.*, **118**, 11–21.

Kim,J.L., Nikolov,D.B. and Burley,S.K. (1993) *Nature*, **365**, 520–527.

Kim,Y., Geiger,J.H., Hahn,S. and Sigler,P.B. (1993) *Nature*, **365**, 512–520.

Kohtz,J.D., Jamison,S.F., Will,C.L., Zuo,P., Lührmann,R., Garcia-Blanco,M.A. and Manley,J.L. (1994) *Nature*, **368**, 119–124.

Krol,A., Westhof,E., Bach,M., Lührmann,R., Ebel,J.-P. and Carbon,P. (1990) *Nucleic Acids Res.*, **18**, 3803–3811.

Lutz,C.S. and Alwine,J.C. (1994) *Genes Dev.*, **8**, 576–586.

Mattaj,I.W. (1993) *Cell*, **73**, 837–840.

Messerle,B.A., Wider,G., Otting,G., Weber,C. and Wüthrich,K. (1989) *J. Magn. Reson.*, **85**, 608–613.

Milligan,J.F., Groebe,D.R., Witherell,G.W. and Uhlenbeck,O.C. (1987) *Nucleic Acids Res.*, **15**, 8783–8789.

Nagai,K., Oubridge,C., Jessen,T.H., Li,J. and Evans,P.R. (1990) *Nature*, **348**, 515–520.

Otting,G. and Wüthrich,K. (1990) *Q. Rev. Biophys.*, **23**, 39–96.

Plateau,P. and Guéron,M. (1982) *J. Am. Chem. Soc.*, **104**, 7310–7311.

Puglisi,J.D. and Tinoco,I.,Jr (1990) *Methods Enzymol.*, **180**, 304–325.

Puglisi,J.D., Tan,R., Canlan,B.J., Frankel,A.D. and Williamson,J.R. (1992) *Science*, **257**, 76–80.

Puglisi,J.D., Chen,L., Frankel,A.D. and Williamson,J.R. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 3680–3684.

Qian,Y.Q., Otting,G., Billeter,M., Müller,M., Gehring,W. and Wüthrich,K. (1993) *J. Mol. Biol.*, **234**, 1070–1083.

Rosbash,M. and Séraphin,B. (1991) *Trends Biochem. Sci.*, **16**, 187–190.

Rould,M.A., Perona,J.J., Soll,D. and Steitz,T.A. (1989) *Science*, **246**, 1135–1142.

Scherly,D., Boelens,W., Dathan,N.A., van Venrooij,W.J. and Mattaj,I.W. (1990) *Nature*, **345**, 502–506.

Scherly,D., Kambach,C., Boelens,W., van Venrooij,W.J. and Mattaj,I.W. (1991) *J. Mol. Biol.*, **219**, 577–584.

Spolar,R.S. and Record,M.T.J. (1994) *Science*, **263**, 777–784.

Steitz,T.A. (1990) *Q. Rev. Biophys.*, **23**, 205–280.

Terns,M.P., Lund,E. and Dahlberg,J.E. (1993) *Nucleic Acids Res.*, **21**, 4569–4573.

van Gelder,C.W.G., Gunderson,S.I., Jansen,E.J.R., Boelens,W.C., Polycarpou-Schwartz,M., Mattaj,I.W. and van Venrooij,W.J. (1993) *EMBO J.*, **12**, 5191–5200.

Varani,G. and Tinoco,I.Jr (1991) *Q. Rev. Biophys.*, **24**, 479–532.

Weeks,K.M. and Crothers,D.M. (1991) *Cell*, **66**, 577–588.

Wimberly,B., Varani,G. and Tinoco,I.,Jr (1993) *Biochemistry*, **32**, 1078–1087.

Wishart,D.S., Sykes,B.D. and Richards,F.M. (1992) *Biochemistry*, **31**, 1647–1651.

Wittekind,M., Görlach,M., Friedrichs,M., Dreyfuss,G. and Mueller,L. (1992) *Biochemistry*, **31**, 6254–6265.

Wüthrich,K. (1986) *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York.

Received on April 21, 1994; revised on May 30, 1994