Journal of Biomolecular NMR, **11:** 59–84, 1998.
KLUWER/ESCOM

Determination of the NMR structure of the complex between U1A protein and its RNA polyadenylation inhibition element

Peter W.A. Howe*, Frédéric H.-T. Allain**, Gabriele Varani & David Neuhaus***
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

Received 16 June 1997; Accepted 31 July 1997

Key words: RNP domain, RNA–protein interactions, U1A protein

Abstract

RNA–protein recognition is critical to post-transcriptional regulation of gene expression, yet poorly understood at the molecular level. The relatively slow progress in understanding this important area of molecular biology is due to difficulties in obtaining good-quality crystals and derivatives, and in preparing samples suitable for NMR investigation. The determination of the structure of the complex between the human U1A protein and its polyadenylation inhibition element is described here. In this paper, we describe the sample preparation, spectral assignments, construction of the NOE-based distance constraints and methodology for calculating the structure of the complex. The structure was determined to an overall precision of 2.03 Å (for all ordered regions), and 1.08 Å for the protein–RNA interface. The patterns of hydrogen bonding and hydrophobic interactions at the interface were analysed statistically using the final ensemble of 31 structures.

Abbreviations: hnRNP, heterologous nuclear ribonucleoprotein; mRNA, messenger RNA; RNP, ribonucleoprotein; U1A-102, amino acids 2–102 of human U1A protein containing the mutations Tyr^{31}His and Gln^{36}Arg; U1A-117, amino acids 2–117 of human U1A protein; PIE–RNA, polyadenylation inhibition element RNA. To aid the distinction between protein and RNA in the text, amino acid residues are referred to using their three-letter codes throughout, except in the figures and Tables 1 and 4a (where the single-letter code is used to save space).

Introduction

RNA–protein recognition is a crucial aspect of many biological processes that still remains relatively poorly understood at the structural and thermodynamic level. This is largely because there are still fewer than 10 atomic resolution structures of RNA–protein complexes solved (Nagai, 1996). The majority of these existing structures are of complexes between tRNAs and aminoacyl-tRNA synthetases or elongation factors (Cusack, 1995), and all but two structures are less than 3 years old.

This relatively slow progress in studying RNA–protein recognition is due to several underlying technical factors. Crystallization of RNA–protein complexes involves synthesizing large quantities of a range of RNA sequence variants to screen for highly diffracting crystals (Oubridge et al., 1995; Price et al., 1995). In addition, the structural flexibility of RNA leads to alternative conformers and generally less rigid structures that may not pack in highly ordered crystals (Price et al., 1997). Despite these difficulties, most atomic resolution information on RNA–protein complexes has been obtained by crystallography. NMR spectroscopic studies of RNA–protein complexes have generally been limited to a qualitative description of sites of interactions through mapping of chemical shift changes upon complex formation (Görlach et al., 1992; Hall, 1994; Howe et al., 1994; Kanaar et al., 1995); using NMR to determine high-resolution structures of

*Present address: Kemisk Institut V, Copenhagen University, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark. **Present address: Department of Chemistry and Biochemistry, University of California, P.O. Box 951569, Los Angeles, CA 90095-1569, U.S.A. ***To whom correspondence should be addressed.

Table 1. List of all 123 intermolecular NOE-derived constraints

RNA	Protein
A24 H8	**S48** HN (3D-s), Hα (3D-s), Hβ (w)
A24 H2	**L44** Hδ (3D-w)
A24 H1′	**L48** HN (3D-w), Hα (3D-w), Hβ (3D-s)
A24 H2″	**S48** Hβ (m)
A24 H3′	**S48** Hβ (m)
G25 H8	**L49** HN (w), Hα (m), Hβ (m), Hγ (3D-s), Hδ (w)
G25 H1′	**S48** Hα (3D-w), **L49** HN (3D-s), Hα (3D-w), Hβ (w), Hγ (3D-s), Hδ (m)
G25 H2″	**L49** HN (3D-s), Hα (3D-w), Hβ (3D-s), Hδ (3D-w)
A39 H2	**L49** HN (3D-w), Hα (3D-s), Hβ (3D-w), Hγ (3D-s), Hδ (w), **R52** Hδ (3D-w), Hϵ (3D-s)
U40 H3	**L49** Hδ (3D-w), **R52** Hy (3D-s), Hδ (3D-s), Hϵ (3D-w)
U40 H1′	**L49** HN (3D-w), Hγ (3D-w), Hδ (m), **R52** Hδ (3D-s)
U40 H2″	**L49** Hγ (3D-w), Hδ (3D-w)
U41 H3	**N16** HN (3D-s), Hα (w), Hβ (3D-s), **K80** Hϵ (3D-s), Hζ (3D-s)
U41 H1′	**N15** Hδ22 (w)
G42 H8	**Y13** OH (w), **N15** Hδ22 (3D-s), **K50** Hα (3D-w)
G42 H1	**G53** HN (3D-s), Hα (3D-s)
G42 H1′	**K50** HN (3D-w), Hα (w), Hβ (3D-s)
G42 H2″	**Y13** OH (w), **K50** Hα (3D-w), Hβ (3D-w)
C43 H6	**Y13** Hϵ (3D-s), OH (3D-w), **M51** Hϵ (3D-w), **K88** Hγ (3D-w), Hδ (3D-s), Hϵ (3D-s)
C43 H5	**Y13** Hβ (3D-w), Hδ (3D-s), Hϵ (3D-s), **K88** Hγ (3D-s), Hδ (3D-s), Hϵ (3D-s), **Q85** Hϵ22 (3D-s)
C43 H1′	**M51** Hϵ (3D-s), **F56** Hϵ (m), Hϵ (m), **A87** Hβ (3D-s), **K88** Hγ (3D-s), Hδ (3D-s), Hϵ (3D-s)
C43 H2″	**Y13** Hϵ (3D-w), **M51** Hϵ (3D-w), **K88** Hϵ (3D-s)
C43 H4′	**M51** Hϵ (3D-s), **F56** Hϵ (3D-w), Hζ (3D-w), **K88** Hδ (3D-w) Hϵ (3D-w)
C43 H5′/H5″	**Y13** Hϵ (3D-w)
A44 H8	**L44** Hδ (3D-w), **M51** Hϵ (3D-s), **F56** Hϵ (3D-s), Hζ (m)
A44 H2	**T11** Hγ (3D-s), **L44** Hα (3D-w), Hβ (3D-s), Hγ (3D-s), Hδ (w), **F56** Hβ (3D-s), **S91** Hα (3D-s), Hβ (3D-w), **D92** HN (3D-w)
A44 H1′	**L44** Hδ (ED-s), **M51** Hϵ (m), **F56** Hδ (3D-s), Hϵ (m), Hζ (m)
A44 H2″	**M51** Hϵ (3D-s)
A44 H3′	**L44** Hδ (3D-w), **M51** Hϵ (3D-w)
A44 H4′	**M51** Hϵ (3D-s)
C45 H6	**L44** Hδ (3D-w)
C45 H5	**L44** Hδ (3D-w), **D92** HN (3D-s)
C45 amino	**T89** Hγ (w), **S91** Hα (3D-s), **D92** HN (ED-s)
C45 H1′	**T11** Hγ (3D-w), **L44** Hγ (3D-w), Hδ (w), **D92** Hβ (3D-s)
C45 H2″	**L44** Hδ (3D-s), **D92** Hβ (3D-s)
C45 H3′	**L44** Hδ (3D-s)
C46 H5	**L44** Hδ (3D-s)

The category of the distance constraint is indicated in parentheses: m (medium; 0–2.9 Å), w (weak; 0–3.5 Å), 3D-s (0–5 Å) and 3D-w (0–7 Å). All methylene and isopropyl groups in the complex were treated as equivalent groups when defining constraints, since no stereoassignments were available.

protein–RNA complexes is especially challenging because of their limited solubility and high molecular weight (>20 kDa).

The development of effective techniques for the preparation of isotopically labelled RNA has made it possible to use NMR to study RNAs of much larger size than was previously possible (Varani et al., 1996), and to obtain structures of significantly improved accuracy and precision (Allain and Varani, 1997). This recent progress has led to the determination of the structure of several complexes involving peptide models derived from regulatory proteins of immunodeficiency viruses (Puglisi et al., 1995; Ye et al., 1995, 1996; Battiste et al., 1996) and to the structure of the complex between the complete RNA-binding domain of the human U1A protein and an internal loop RNA target (Allain et al., 1996).

The human U1A protein represents a paradigm for understanding RNA recognition by proteins of the RNA-processing machinery. The determination of the crystal structure of the complex between U1A and a hairpin RNA substrate (Oubridge et al., 1994) and the NMR structure of the U1A–internal loop complex (Allain et al., 1996) have provided an unprecedented insight into structural and dynamic aspects of RNA–protein recognition. Direct comparison of the complex with the solution structures of the uncomplexed RNA (Gubser and Varani, 1996) and protein components (Avis et al., 1996) have led to the proposal of a mechanism for binding (Allain et al., 1996). In this report, we describe in detail the determination of the NMR structure of the complex between the human U1A protein RNA-binding domain of 102 amino acids and part of the polyadenylation inhibition element RNA from the 3′-untranslated region of the U1A pre-mRNA (also called PIE-RNA (Gunderson et al., 1997)). Completion of this task involved solving several technical problems in sample preparation, spectral assignments and construction of the list of NOE-based distance constraints. It was also necessary to develop methodology for calculating the structure of the complex; a molecular dynamics protocol was used starting directly from random RNA and protein initial coordinates without using any *ad hoc* assumptions or docking steps.

Materials and methods

Sample preparation

Two protein constructs were used in this study. The first corresponds to residues 2–102 from human U1A protein (U1A-102), and the second contains two point mutations, $Tyr^{31} \rightarrow$His and $Gln^{36} \rightarrow$Arg (see below). Both constructs include the RNA-binding domain of U1A (residues 5–98) and both bind RNA as tightly as the full-length, wild-type U1A protein (Nagai et al., 1990; Gubser and Varani, 1996). Both were overexpressed in BL21(DE3) *E. coli*, the wild-type using the plasmid pMW172 and the double-mutant using a plasmid based on pET13a (Gerchman et al., 1994). The pET13a vector carries a kanamycin-resistance gene and a strictly regulated promoter, to provide reproducibly high levels of protein expression. Bacteria were grown on M9 minimal medium supplemented by thiamine and trace elements, and stable isotopes were incorporated into the protein by using [U]-$^{13}C_6$-glucose and [^{15}N]-ammonium chloride as required.

Protein was purified essentially as described previously (Nagai et al., 1990), except for some minor modifications. After purification, protein was stored frozen at -70 °C in 10 mM KH_2PO_4, 25 mM KCl at pH 7.4. NMR samples of the free proteins contained 1–2 mM protein at pH 4.9 in 10 mM KH_2PO_4, 10 mM [$^2H_4$]-sodium acetate and 0.02% azide. Samples were prepared either in 90% H_2O / 10% D_2O or in 100% D_2O, and NMR data were collected at 300, 308 and 315 K.

The RNA was prepared by *in vitro* transcription from synthetic DNA templates using T7 polymerase as previously described (Gubser and Varani, 1996). RNA was purified by denaturing polyacrylamide gel electrophoresis, followed by ethanol precipitation, extensive dialysis and size exclusion chromatography (Varani et al., 1996). The sequence of the RNA construct used for this study is shown in Figure 1. Substitution of an exceptionally stable UUCG tetraloop in place of the wild-type loop sequence improved thermodynamic stability and simplified assignments without altering the affinity of the RNA for the protein (Gubser and Varani, 1996).

Preparation of the protein–RNA complexes

NMR samples were prepared by dialysing separately protein and RNA into 10 mM KH_2PO_4 with 0.02% sodium azide (pH $\approx$ 5.0). For samples in 100% D_2O, the components were freeze-dried after dialysis and separately resuspended in D_2O. For each of

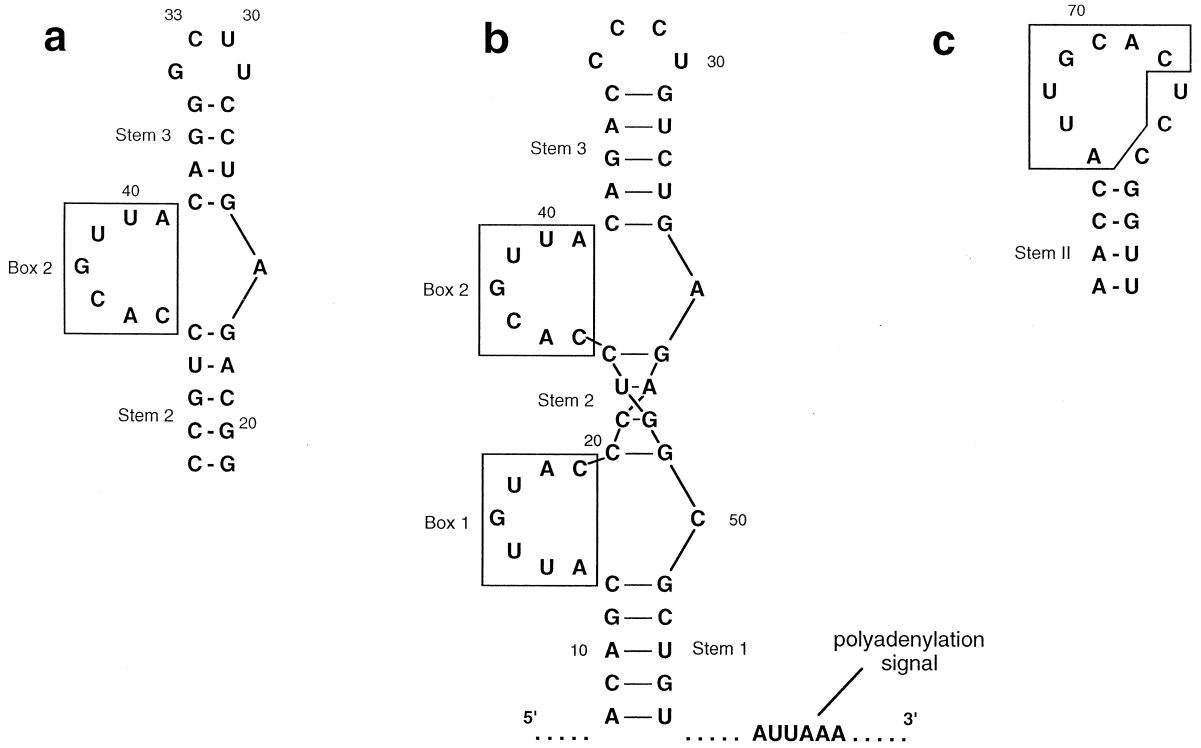


Figure 1. Secondary structures of the U1A 3′UTR PIE (a) box 2 from the present study and (b) the complete PIE-RNA. In (c) is shown the secondary structure of stem-loop II of U1 snRNA; the boxed sequences are common to all three structures, except for a C → U mutation in box 1 of the complete PIE-RNA.

the components, concentrations were measured by UV absorbance, the pH was adjusted to 4.9 by adding [$^2H_4$]-sodium acetate to 10 mM and 10% D_2O was added to samples in water. The complex was then formed by gradually adding protein to the RNA at room temperature. Samples became very cloudy on initial addition of protein, but rapidly cleared upon gentle mixing; they were then spun through a 0.2 μm spin-filter (Amicon Inc., Beverly, MA) to remove any particles and to sterilize the samples. RNA was always kept in slight (5–10%) excess; if excess protein was inadvertently added, some precipitation was observed. Samples contained complex at about 1 mM concentration, and susceptibility-matched, variable-restricted volume NMR tubes (Shigemi Inc., Tokyo) were used so as to minimize the amount of material needed. Although it is conceivable that the cytosine bases of the RNA might be protonated at pH 4.9 (the pK for N3 of cytosine is approximately 4.5), we believe no such complication occurred in these studies, for the following reasons: (i) the NMR spectra were largely unchanged over the range pH 5–7; (ii) guanines paired with cytosines in the stems (C21, C27, C28, C38, C46, C49 and C50) gave imino signals characteristic of normal Watson–Crick base pairing, inconsistent with cytosine protonation; (iii) analysis of the structures strongly suggests that the N3 atoms of both C43 and C45 act as hydrogen-bond acceptors in hydrogen bonds that are also found in the U1A/stem-loop II crystal structure (cf. Table 4b); and (iv) no exchangeable signals were observed near 15 ppm, as would have been expected for protons attached to cytosine N3. All NMR spectra used to make assignments for the complex were recorded at 300 K.

Improving the stability of the protein–RNA complex

Preliminary work on the complex between wild-type U1A-102 and RNA showed that the complex slowly precipitated over several days at the concentrations and temperatures necessary to obtain acceptably narrow NMR resonances (>20 °C). Screening of a broad range of conditions of pH, solvent composition and ionic strength did not prevent this, although low ionic strength did improve sample behaviour considerably. The use of a doubly mutated U1A protein originally developed for crystallization of the complex with the

hairpin loop RNA substrate (Oubridge et al., 1994, 1995) improved the long-term stability of the NMR sample. The two mutations, $Tyr^{31} \rightarrow$ His and $Gln^{36} \rightarrow$ Arg, change exposed hydrophobic residues from helix A into charged residues. The dissociation constant of the complex is not affected by these mutations (Oubridge et al., 1994, 1995; Gubser and Varani, 1996) and the chemical shifts of resonances from Tyr^{31} and Gln^{36} are not perturbed upon RNA binding (Howe et al., 1994). Protein purity, and hence probably also stability of the complex, was improved by limiting the portion retained during chromatographic separations. Complexes prepared with the double-mutant protein were stable for several weeks at temperatures of 27 °C; however, even with the double-mutant protein, some samples of complex deteriorated quickly due to RNA hydrolysis, presumably caused by RNAse contaminants in the protein preparation.

NMR spectroscopy

All data were acquired using either Bruker DMX600 or Bruker AMX500 NMR spectrometers. The 600 MHz spectrometer was equipped with a 5 mm triple-resonance $^1H\{^{13}C$, broadband$\}$ probe with z-axis pulsed field gradient coil, and the 500 MHz spectrometer was equipped with a 5 mm double-resonance $^1H\{$broadband$\}$ probe or a 5 mm triple-resonance $^1H\{^{13}C,^{15}N\}$ probe.

Free protein resonances were assigned using a variety of spectra, including 2D $[^1H,^1H]$ experiments (TOCSY, NOESY, DQ correlation and RELAY) and 2D and 3D $[^1H,^{15}N]$ experiments with ^{15}N-labelled protein (2D-$[^{15}N]$-HSQC, 2D-$[^{15}N]$-HSQC-TOCSY and 3D-NOESY-$[^{15}N]$-HSQC). An HNCACB experiment (Wittekind and Mueller, 1993) was recorded using double-labelled protein in water, and various other experiments were recorded using double-labelled protein in D_2O ($[^{13}C]$-HMQC, 2D-$[^{13}C]$-HSQC-TOCSY, constant time 2D-$[^{13}C]$-HSQC-TOCSY, 2D-$[^{13}C,^1H]$-HCCH-COSY and 3D-NOESY-$[^{13}C]$-HMQC). $[^1H,^1H]$ isotropic mixing in TOCSY experiments was achieved using the DIPSI-2 sequence for 40 or 80 ms (Shaka et al., 1988), and NOESY experiments employed a mixing time of 200 ms. GARP-1 decoupling of ^{15}N and ^{13}C was used during the acquisition time as necessary (Shaka et al., 1985).

For the complex, samples in H_2O containing ^{15}N-labelled protein and unlabelled RNA were used to obtain homonuclear NOESY and DQ spectra, as well as $[^{15}N]$-HSQC and 3D-NOESY-$[^{15}N]$-HMQC spectra. HNCACB, HN(CO)CA, HBHA(CBCACO)NH, 3D-NOESY-$[^{15}N]$-HSQC and 3D-$[^{13}C]$-HSQC-NOESY spectra were recorded from samples in H_2O containing double-labelled protein and unlabelled RNA, and similar samples of complex in D_2O were used to record $[^{13}C]$-HSQC, 3D-HCCH-TOCSY, 3D-HCCH-COSY, 2D-HCCH-COSY (optimized for aromatic resonances), 3D-NOESY-$[^{13}C]$-HMQC spectra and half-filtered 2D $[^1H,^1H]$ NOESY experiments (Otting and Wüthrich, 1990; Clore and Gronenborn, 1994). The HCCH-TOCSY experiment employed a 17 ms DIPSI-2 isotropic mixing sequence, while NOESY experiments employed mixing times of 50 or 100 ms.

The spectral width for ^1H was typically set to 8 kHz, except for the indirect dimension in DQ correlation spectra (16 kHz) and the indirect ^1H dimensions of 3D experiments recorded in D_2O (4 kHz). ^{15}N spectral widths were 32.89 ppm and the ^{13}C spectral width was 66.26 ppm. Two-dimensional experiments for the free protein typically used 1024 complex points in F2 and 512 real points in F1. All the other experiments used 512 complex points in the acquisition dimension and 256 real points in indirect ^1H dimensions. Heteronuclear dimensions used either 64 (3D experiments) or 256 (2D experiments) real points. TPPI was used for sign discrimination in indirect dimensions (Marion and Wüthrich, 1983), except for gradient-selected and sensitivity-enhanced gradient-selected experiments, which used 'echo anti-echo' sign discrimination (Davis et al., 1992). Initial delays for indirect dimensions were set to multiples of the incrementation time to improve baselines and give defined phases for folded signals (Bax et al., 1991).

In most homonuclear experiments, the residual solvent signal was suppressed by coherent on-resonance presaturation, while in most heteronuclear experiments it was suppressed using spin-lock rf purge pulses supplemented by z-gradients. In some NOESY spectra, jump-return excitation was used to help preserve the intensity of NOE cross peaks by avoiding solvent saturation. In most 2D-HSQC and HMQC experiments, z-gradients used for coherence selection also suppressed the solvent signal.

Many 2D experiments were processed using UXNMR (Bruker, Karlsruhe) running on either an AspectStation1 or an X32 computer. Other experiments were processed using the program Felix (Biosym Inc., San Diego, CA) running on Silicon Graphics Indigo or Indy workstations. Before transformation, time-domain data were multiplied by sine or sine-squared functions shifted by 30–60° and zero-filled

twice (2D spectra) or once (3D spectra) in each dimension. Three-dimensional spectra recorded in water were baseline corrected using time-domain convolution (Marion et al., 1989; Waltho and Cavanagh, 1993).

Spectral assignments

The assignment of the RNA component, both free and in the complex, has been described previously (Gubser and Varani, 1996); only the assignments of the protein component are discussed here.

Proton and ^{15}N assignments for free, wild-type U1A-102 protein, and also for the double-mutant Tyr^{31}His Gln^{36}Arg protein, were obtained using standard techniques applied to both unlabelled and ^{15}N-labelled protein samples (Wüthrich, 1986). Amino acid spin systems were identified and classified according to residue type using homonuclear TOCSY, COSY and RELAY spectra, together with a 2D-HSQC-TOCSY spectrum. Spin systems were arranged within the known primary sequence using sequential NOE cross peaks in 2D-HSQC-NOESY and 3D-NOESY-HSQC spectra. This process was facilitated by the availability of previously published partial backbone proton assignments for a fragment of U1A containing residues 11–94 (however, the shifts previously reported for Phe^{75} and Lys^{80} we believe to be due to Lys^{80} and Asp^{79} respectively) (Hoffman et al., 1991). Assignments were generally straightforward to obtain, except for two regions (Ser^{46}–Leu^{49} and Asp^{90}–Asp^{92}) where amide NH resonances were not detectable; in the case of Asp^{90}, none of the resonances were observed. However, these regions are relatively short and contain characteristic spin systems, so it was possible to assign the other resonances from these residues by a process of elimination.

In the case of the Tyr^{31}His, Gln^{36}Arg double-mutant protein, a series of triple-resonance experiments was recorded using [^{15}N, ^{13}C] double-labelled samples to generate the carbon assignments. An HN-CACB spectrum was used to identify α and β carbon shifts and to verify the sequential assignment, while the remaining carbon shifts were identified using [^{13}C]-HMQC, 2D-HCCH-COSY, 3D-NOESY-[^{13}C]-HMQC and 2D-[^{13}C]-HSQC-TOCSY spectra. A constant-time 2D-[^{13}C]-HSQC-TOCSY showed few correlations, but the higher resolution resolved ambiguities in the crowded methyl region, and a constant-time [^{13}C]-HSQC acquired without carbonyl decoupling allowed unambiguous identification of the 10 serine CβH$_2$ groups.

The strategy for assigning the spectra of the U1A102–RNA complex was essentially to build upon the results already obtained for the free components. As expected, spectra from the free components were of significantly higher quality than those from the complex, showing more complete connectivity patterns, particularly in experiments based on scalar couplings. Although it might have been possible to obtain assignments using only NMR spectra from the complex, the combined analysis of data from the complex and from both free components led to a much more secure and rapid interpretation.

Many protein resonances in the complex gave chemical shifts and patterns of NOE cross peaks similar to those observed for the free protein. Previous NMR and crystallographic studies of RNP-RNA complexes have shown that no *widespread* change in protein conformation occurs upon RNA binding (Görlach et al., 1992; Howe et al., 1994; Oubridge et al., 1994), so we reasoned that most resonances of residues not directly involved in RNA binding would be largely unperturbed upon complexation. Initially, comparisons were made in the 3D-NOESY-[^{15}N]-HMQC and 3D-NOESY-[^{15}N]-HSQC spectra of free and complexed protein; if an NH correlation was found at similar locations in corresponding HMQC planes from the two spectra, and if the pattern of NOE cross peaks observed in the (orthogonal) NOE dimension starting from this NH correlation was also similar in the two spectra, then it was tentatively concluded that the assignment for this NH group was the same in the complex as for the free protein. Such tentative assignments were then checked and extended by searching specifically for NOE cross peaks to sequentially neighbouring spin systems in spectra of the complex. By this process, groups of sequentially neighbouring spin systems were constructed, including whole sets for each of which the pattern of NOE cross peaks matched between the spectra of free and bound protein. Initially, only resonances with very closely similar chemical shifts and patterns of NOE cross peaks were accepted (as shown in Figure 2 for Asn^{67}), but as the number of unassigned NH signals diminished, it became possible to assign securely resonances which had undergone larger changes in chemical shift upon RNA binding. The distinction between protein and RNA signals in the pattern of NOE cross peaks observed from each protein NH group was made partly on grounds of chemical shift (RNA generally gives no signals below 3.5 ppm), and also by checking

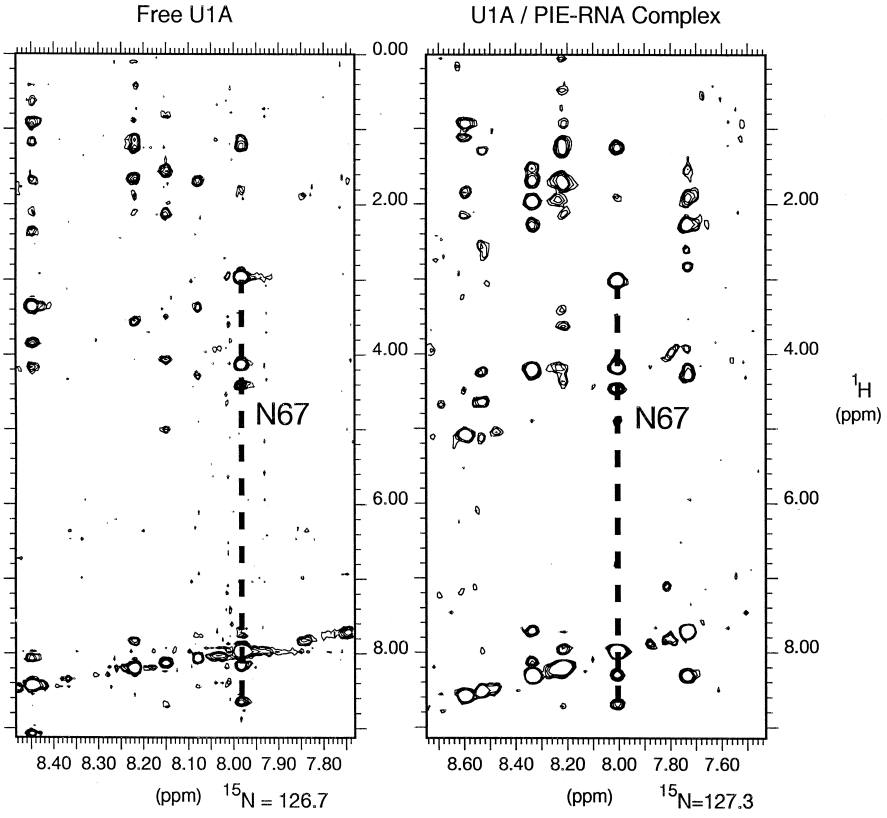


Figure 2. F2-planes from 3D-[^{15}N]-NOESY-HSQC spectra of ^{15}N-labelled protein free (left) and bound to PIE-RNA (right). In each case the plane corresponds to the ^{15}N shift of residue Asn^{67}, and intraresidual correlations from the ^{1}H shift of Asn^{67} NH are indicated by a vertical dotted line.

data from [^{1}H,^{13}C] spectra acquired from complex containing double-labelled protein.

This comparative method of assignment was particularly effective for the helical regions of the protein, because relatively strong sequential NH-NH NOE cross peaks could be used to link together spin systems of adjacent residues. Since the first two helices of the protein (residues 24–35 and 63–71) are distant from the RNA binding site, they undergo very small changes in chemical shift upon RNA binding and were quickly assigned. Residues in helix C were somewhat more difficult to assign because they undergo large chemical shift changes upon complexation. By including long-range NOE cross peaks in the analysis (NOE cross peaks between β-strands), it was also possible to assign some parts of the β-sheets despite the larger chemical shift changes in this region of the protein.

The remainder of the RNA-bound protein (about 30 residues) could not be assigned by comparison with the free state, because large changes in chemical shift occurred on complexation. For these residues, it was necessary to use data from complexes containing [^{13}C,^{15}N]-double-labelled protein to obtain heteronuclear through-bond correlations and ^{13}C-chemical shifts. A range of through-bond experiments were recorded (Figure 3); HCCH-COSY and HCCH-TOCSY experiments allowed identification of some spin systems, particularly from residues containing methyl groups. A homonuclear double-quantum experiment helped in the assignment of aromatic spin systems and also provided about 30 useful intraresidual NH-Hα correlations. However, most spin systems remained incomplete and some residues gave no correlations at all in through-bond experiments (Figure 3). Thus, the assignment was completed using 3D-NOESY-[^{15}N]-HSQC and 3D-[^{13}C]-HSQC-NOESY spectra (all recorded using the same double-labelled sample, to eliminate small chemical shift differences between different preparations). The 3D-NOESY spectra allowed assignment of most of the ^{13}C chemical shifts of the protein in the complex, so that the approach of assignment by comparison to the

free protein could be extended through the remaining regions of the protein. ^{13}C chemical shifts were more similar between free protein and complex than were either ^{1}H or ^{15}N shifts, and, as expected, ^{13}C chemical shifts also provided a more reliable guide to residue type identification (Grzesiek and Bax, 1993). For many residues, results from backbone experiments allowed inter- and intraresidue NOE cross peaks to be differentiated. This combination of NOE pattern comparison, through-bond experiments and ^{13}C chemical shift analysis allowed a nearly complete assignment of the protein in the complex.

Intramolecular distance constraints

The NOE-derived intramolecular distance constraints for the RNA-bound U1A-102 protein were obtained using the constraints for the free U1A-117 protein as a starting point. Of the 1710 NOE-derived intramolecular distance constraints identified for the complex, 1287 (75%) are found in both free and bound protein spectra. Most differences between the two constraint lists arise from differences in the dispersion of resonances in the two sets of spectra, from the two mutations, and from the fact that stereoassignments obtained for the free protein could not be obtained for the complex. Other differences have genuine structural implications (see below). In addition to the distance constraints, 42 hydrogen-bonding constraints (two constraints per hydrogen bond) were obtained upon the observation of 21 slowly exchanging amide signals in spectra of the complex (data not shown).

As with the protein, the constraint list for the RNA in the complex was constructed using the list for the free RNA as a template (Gubser and Varani, 1996). Of the 591 intramolecular NOE-derived constraints for the RNA in the complex, 526 (89%) were also present in the free RNA constraint list, as were the 25 hydrogen-bonding constraints. In addition, 110 dihedral constraints were introduced for the complex, using the same constraints derived experimentally for the free RNA in all regions of the double-helical stems where no chemical shift differences were observed between free and bound RNA. Of the 65 constraints present only for the bound RNA, 32 reflect genuine structural differences (see below).

Identification of intermolecular NOE interactions

Two-dimensional half-filter experiments were recorded to observe selectively intermolecular connectivities from samples containing isotopically labelled protein in the presence of unlabelled RNA (Otting and Wüthrich, 1990). These spectra contain some very intense peaks corresponding to intermolecular NOE connectivities, which were relatively simple to assign. However, owing to spectral overlap and artefacts, the half-filter spectra were not of sufficient quality to assign all the intermolecular NOE interactions. Most intermolecular NOE cross peaks were identified from 3D ^{13}C-edited spectra acquired both in H_2O and in D_2O with labelled protein complexed to unlabelled RNA (Figure 4). Intermolecular NOE cross peaks involving ^{15}N-bound protein NH signals were observed in the 3D ^{15}N-edited spectrum, and were assigned as intermolecular by establishing whether the corresponding cross peaks were present or absent in the 3D ^{13}C-edited NOESY spectrum recorded for the same sample in H_2O. Since intermolecular interactions would link NH protons on the protein to ^{12}C-bound protons on the RNA, they should be missing in the latter spectrum. Some intermolecular contacts could only be identified in 2D-NOESY spectra (D_2O and H_2O), but these cases required particular care since inter- and intramolecular connectivities can only be distinguished in such spectra when the chemical shifts involved are unique. Nonetheless, 2D-NOESY spectra were essential to assign intermolecular NOE interactions involving the RNA imino and amino protons and the Tyr^{13} and Phe^{56} aromatic protons.

As described above, approximately 30 strong and unambiguous intermolecular NOE connectivities could be obtained straightforwardly; this set included mainly H1′ and aromatic RNA resonances and methyl and amide protein resonances (Figures 4a and b). These 30 intermolecular connectivities provided the starting point for assigning the remaining contacts (Table 1). A second set of ≈30 intense NOE cross peaks could be assigned based on these, since they were closely related to resonances involved in the first set. Typically, these included connectivities from a protein side-chain methylene (Hβ or Hγ) and were assigned because of unambiguous intermolecular NOE cross peaks involving well-dispersed methyl or amide resonances of the same protein residue. A third set of intermolecular NOE cross peaks (≈ 30) included some that were very weak, just above the noise level, which were constrained in the range 0–7 Å. Again, these could only be assigned because of a close relationship with intermolecular NOE cross peaks that had already been assigned, as illustrated in Figures 4c–f. A final set of 35 intermolecular contacts could only be assigned after re-examining NMR spectra using preliminary calculated structures to identify am-

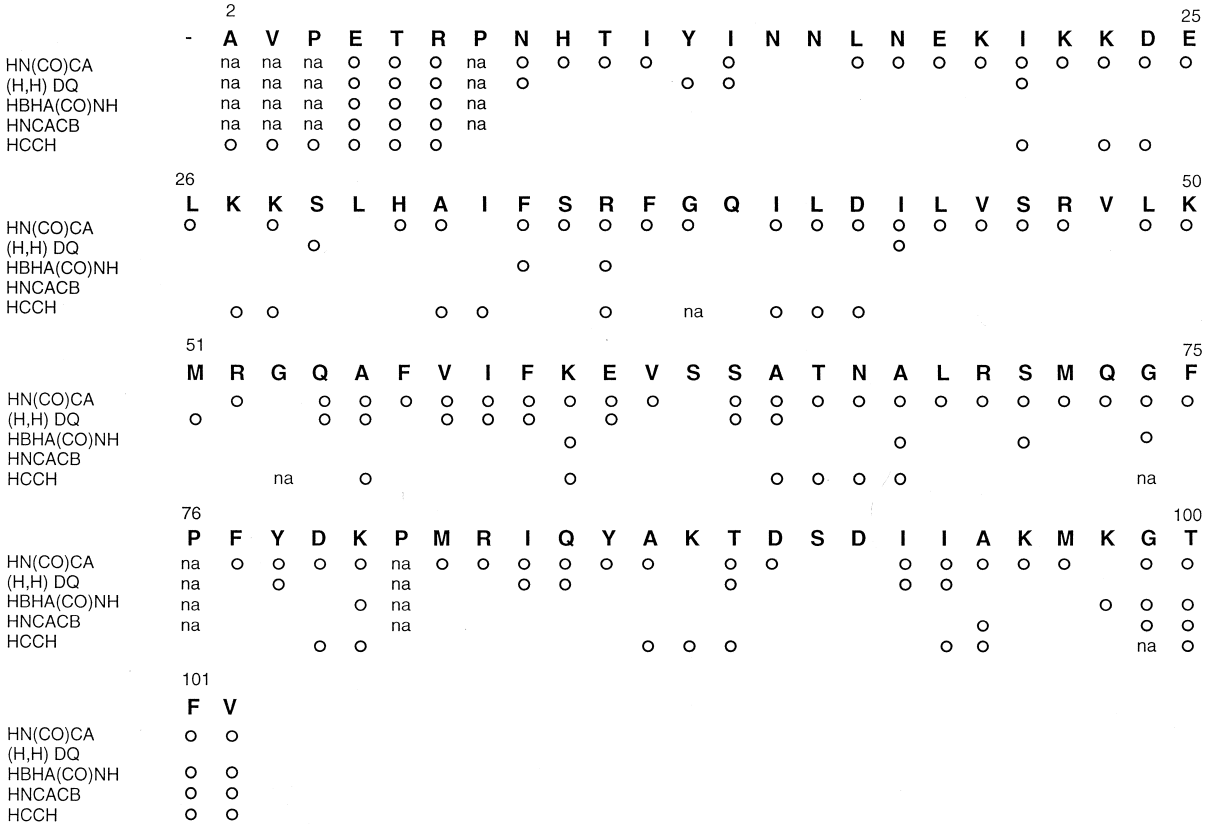


Figure 3. A summary of the results of through-bond correlation experiments with U1A protein complexed to box 2 from the PIE-RNA. Each correlation detected is indicated as ○, (H,H) DQ indicates homonuclear (^{1}H,^{1}H) double-quantum correlation, HCCH indicates either HCCH-COSY or HCCH-TOCSY, and 'na' indicates not applicable.

biguous connectivities. These connectivities involve mostly RNA ribose resonances (H2′, H3′, H4′, H5′ and H5″) that could not be unambiguously assigned using chemical shifts alone due to the poor dispersion of those proton resonances and overlap with Hα protein resonances.

Calibration of the distance constraints

Most intermolecular NOE contacts were calibrated in the same way as intramolecular NOE contacts by using 2D homonuclear NOESY spectra recorded at 50 and 100 ms mixing time and 3D ^{15}N- and ^{13}C-edited NOESY spectra acquired at 100 ms mixing times. Where possible, calibration of intensity was carried out using the 50 ms NOESY data, taking as references the most intense $d_{\alpha N}$ connectivities in the β-sheet regions (2.3 Å) and the $d_{\alpha N}$ (i, i+3) connectivities in the α-helices (3.5 Å). Based on this spectrum, upper bound categories of 'strong' (2.3 Å), 'medium' (2.9 Å) or 'weak' (3.5 Å) were defined. In order to use data from the 100 ms mixing time 3D-NOESY spectra, which included many connectivities obscured by overlap in the homonuclear data, a more conservative approach was required since spin diffusion at longer mixing times distorts the linear relationship with cross-peak intensity. For these data, two additional upper bound categories were therefore defined: '3D strong' (5 Å) and '3D weak' (7 Å), based on relative intensity in the 100 ms data. These longer upper bounds served as a precaution against possible misinterpretation of NOE intensities arising from spin diffusion (Avis et al., 1996). Upper bounds for constraints involving equivalent or non-stereoassigned protons were increased to allow for multiplicity; for methyl groups, the upper bound was increased by $3^{1/6}$ (i.e. approximately 20%), and for degenerate aromatic or CH_2 signals the upper bound was increased by $2^{1/6}$ (i.e. approximately 12%) (Fletcher et al., 1996). Lower bounds for the NOE-derived distance constraints were all set to 0 Å so as not to interfere with the operation of the initial search phase of the sim-

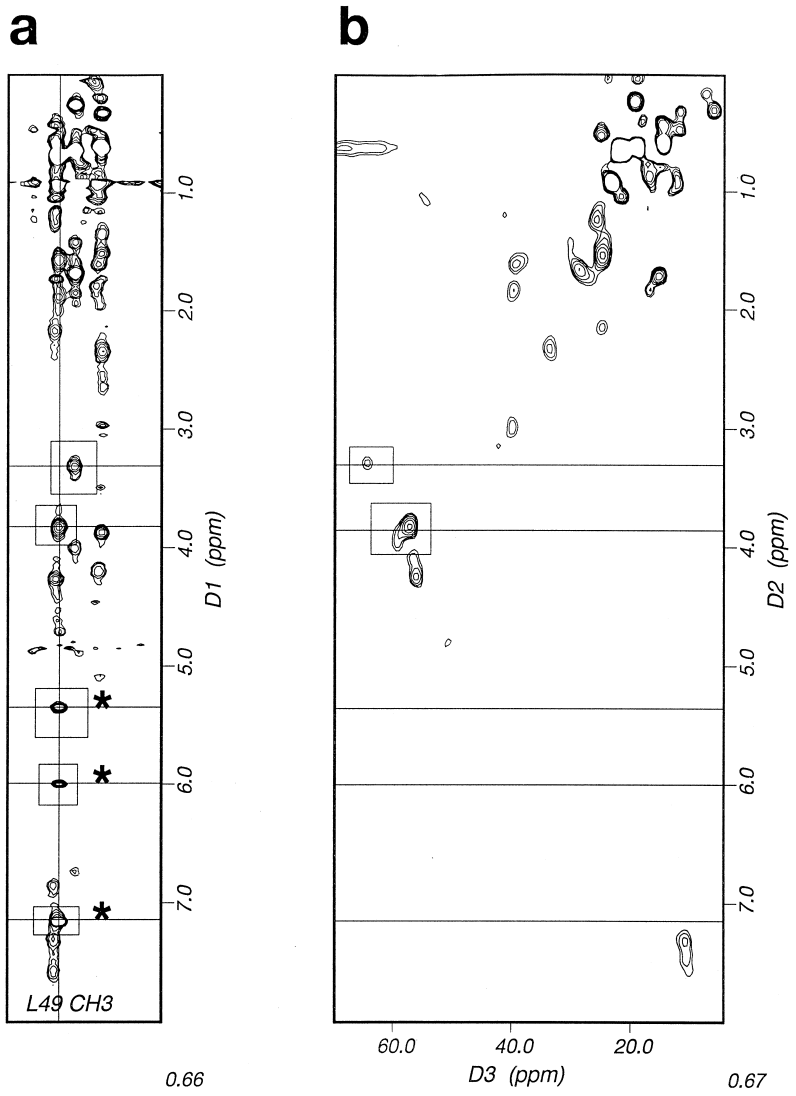


Figure 4. Two-dimensional sections of the 3D ^{13}C-edited NOESY spectrum of the complex in D_2O. For interactions that are sufficiently strong, intermolecular NOE contacts can be identified by comparing a 'NOESY plane' (a) with the corresponding 'HSQC planes' (b). Intermolecular NOE contacts have no counterpart in the HSQC type plane, since the RNA is unlabelled. For example, Leu^{49} CδH$_3$ at 0.66 ppm shows three intermolecular NOE cross peaks between 5 and 7 ppm (spectrum on the left, boxes marked $*$), identifiable by the fact that they have no counterpart in the spectrum on the right. The remaining cross peaks do have counterparts in the HSQC plane (e.g. the other two boxed cross peaks), and are intramolecular in origin. Once a strong intermolecular interaction has been identified in this way, it is often possible to use this as a basis for other assignments. For instance, comparison of the 2D plane at the chemical shift of Leu^{49} Hδ (panel (c); corresponds to the lower part of panel (a)) with those at the chemical shifts of Leu^{49} Hγ (d), Hβ (e) and Hα (f) shows that the strong peaks from the Hδ protons already identified as being intermolecular have counterparts in the other panels that align exactly and can therefore be assigned by analogy.

ulated annealing protocol, during which the van der Waals radii of all atoms are greatly reduced to allow the polypeptide chain to pass through itself freely, as previously explained by Hommel et al. (1992). Given that van der Waals interactions are handled explicitly (and separately from the NOE-derived constraints) by the simulated annealing protocols (Nilges et al., 1988; Wimberly, 1992; Gubser and Varani, 1996; Varani et al., 1996), this also avoids any unwanted duplication in specifying the van der Waals interactions corresponding to those proton pairs for which there are NOE constraints.

Intermolecular NOE cross peaks are generally broader than most intramolecular NOE cross peaks,

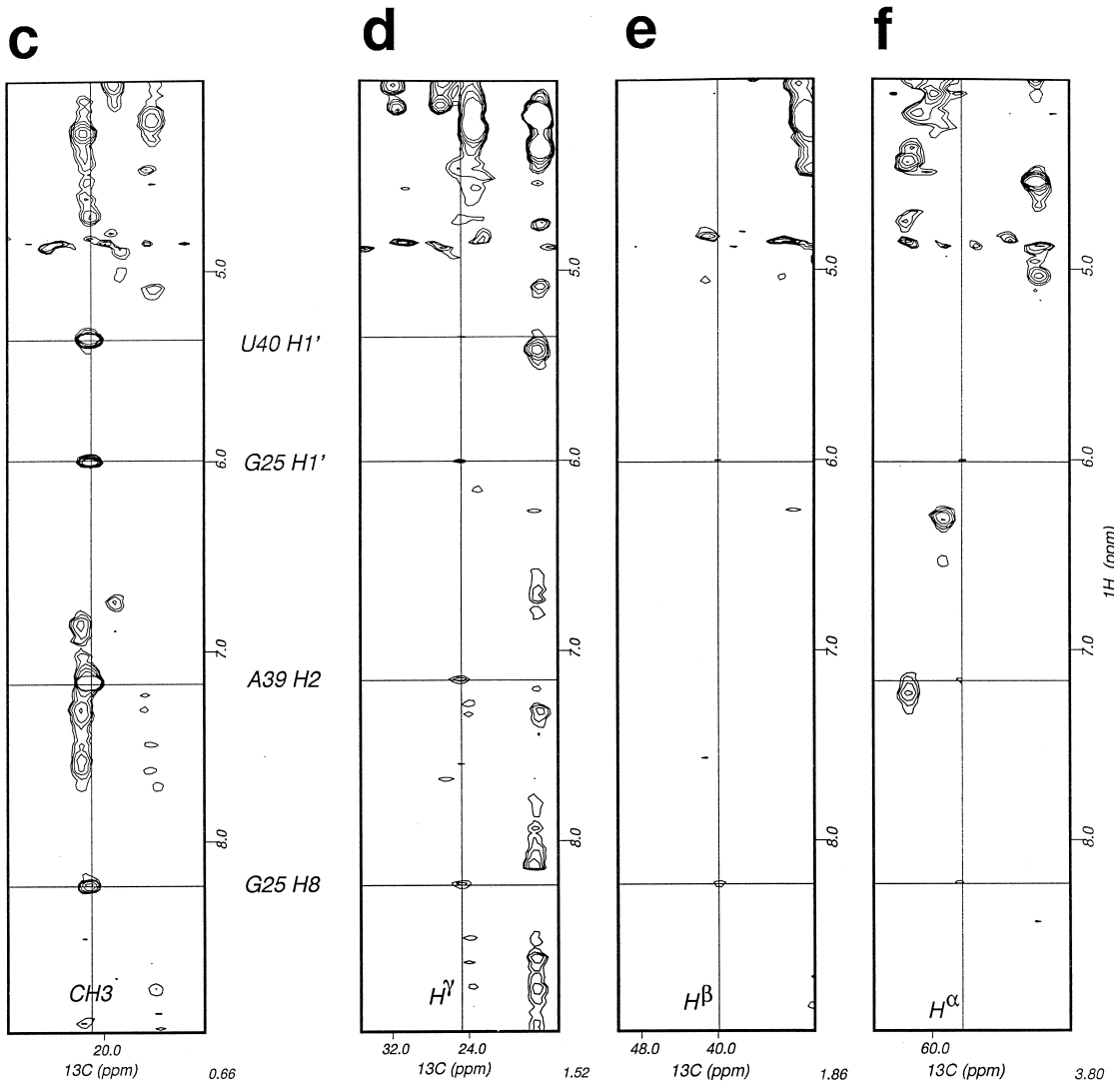


Figure 4. Continued.

possibly due at least in part to local motions at the RNA–protein interface, and are therefore more difficult to observe. This explains why a large number (41 out of 123) of such constraints are in the category 0–7 Å. However, 25 intermolecular NOE contacts were classified as either 0–2.9 Å (medium) or 0–3.5 Å (weak) (Table 1). Half of these constraints were obtained from 2D spectra, where more precise volume integration and calibration were possible, while the other half were only resolved in 3D spectra. These latter cross peaks were classified by comparing their intensities with peaks of known intensity in the same 3D plane (covalently linked protons) for which corresponding interproton distances were known from preliminary rounds of calculated structures.

Structure calculations

Fifty structures of the complex were calculated using a novel restrained molecular dynamics protocol implemented within the program X-PLOR 3.1 (Brünger, 1990). This protocol results from an empirical combination of elements from previously developed protein structure calculation protocols and RNA structure calculation protocols, neither of which were successful when used alone. This protocol is considered further in the Results and Discussion section.

Refinement of the structure

Several rounds of calculations were performed to improve the precision of the structure. First, the protein component (only) in the complex was calculated using 1729 intra-protein distance constraints. At this stage, few violations were detected and the precision (mean rmsd to the mean structure) was 0.92 ± 0.11 Å for the entire protein backbone. Acceptors corresponding to each of the 21 slowly exchanging protein backbone amides were identified unambiguously in this structure, and appropriate hydrogen-bonding constraints were introduced. A first structure of the complex was next calculated, using the combined protocol referred to above with 2546 constraints, including 61 intermolecular distance constraints. The resulting protein backbone was well defined (rmsd 0.74 ± 0.2 Å) for 16 out of 20 structures, but the RNA was poorly defined in the interfacial region; the seven-nucleotide loop (A39–C45) had an rmsd of 2.2 ± 0.4 Å, while the 12 nucleotides of the full RNA binding site (C38–C46 and G23–G25) had an rmsd of 3.2 ± 1.0 Å. Only five intermolecular distance constraints to exchangeable protons were included in this first list of 61 intermolecular distance constraints.

Addition of an extra set of 35 intermolecular NOE-derived constraints (20 of them involving exchangeable resonances) based on further analysis of the spectra in conjunction with the preliminary structures improved the rmsd of the RNA interface to 1.02 ± 0.2 Å for the seven-nucleotide loop and 1.29 ± 0.3 Å for the full RNA binding site. Most of this improvement flowed from the assignment of the imino protons of U40, U41 and G42, which are in relatively slow exchange with solvent and from which 14 NOE connectivities (including 11 intermolecular constraints) were found. This intermediate set of calculated structures was used to identify a final set of 28 intermolecular NOE cross peaks, mostly involving poorly dispersed ribose protons. Contacts to H2′ and H3′ of A24 and H2′ of G25, obtained at this stage, were essential to define precisely the positions of A24 and the two base pairs closing the double-helical stems (G25•C38 and G23•C46). Using this constraint set, the final round of calculated structures yielded an rmsd of 0.87 ± 0.11 Å for the seven-nucleotide loop and 1.01 ± 0.16 Å for the full RNA binding site. The final list contains 2602 constraints as detailed in Table 2; in all, 123 of these are intermolecular constraints, 31 of which involve exchangeable protons.

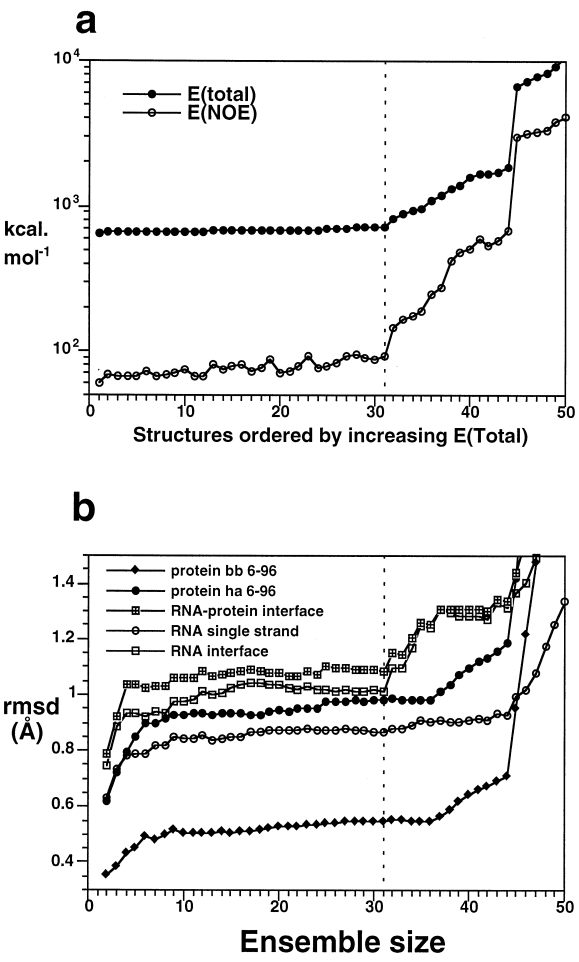


Figure 5. (a) E_{total} and E_{NOE} for the 50 calculated structures, ordered by increasing total energy (E_{total}). The dotted line indicates the 31st structure, which represents the cut-off for the ensemble of converged structures. (b) Rmsd profiles calculated with the program CLUSTERPOSE (Diamond, 1995) for different ensembles of atoms. 'Protein bb' indicates backbone atoms (N, Cα, C′) and 'protein ha' indicates all protein heavy atoms; RNA and RNA–protein calculations are for all heavy atoms.

Analysis of the structure

The same procedure for selection of the 'converged' structures as described previously for RNA (Varani et al., 1996) and protein structures (Avis et al., 1996; Fletcher et al., 1996) was used for the complex at each step of the refinement process. Ensembles were built up by stepwise addition of individual structures in order of increasing total energy; for each successive ensemble size, the average rmsd to the average structure was calculated independently based on global best-fit superposition using the program CLUSTERPOSE (Diamond, 1995). Energy profiles and energy-ordered

Table 2. Distance and angle statistics for the final constraint list

I. U1A–102 protein in complex				
A. NOE-derived distance constraints				
Intraresidue	565	Strong	(0.0–2.3 Å)	14
Sequential	444	Medium	(0.0–2.9 Å)	122
Medium range	252	Weak	(0.0–3.5 Å)	108
Long range	449	3D strong	(0.0–5.0 Å)	982
		3D weak	(0.0–7.0 Å)	484
Total NOE constraints	**1710**			
B. Other constraints				
Dihedral angle (χ^1)	0			
Hydrogen bonds	21 (42 distances for 21 hydrogen bonds)			
II. PIE-RNA (single site, 30 nt) in complex				
A. NOE-derived distance constraints				
Intraresidue	347	Strong	(0.0–2.3 Å)	1
Sequential	175	Medium	(0.0–3.0 Å)	57
Medium range	15	Weak	(0.0–4.0 Å)	131
Long range	54	3D strong	(0.0–5.0 Å)	248
		3D weak	(0.0–7.0 Å)	154
Total NOE constraints	**591**			
B. Other constraints				
Dihedral angle (χ^1)	110			
Hydrogen bond	25 (25 distances for 25 hydrogen bonds)			
III. Intermolecular NOE constraints				
Single-stranded RNA 39–45	98	Strong	(0.0–2.3 Å)	0
A24 bulge	9	Medium	(0.0–2.9 Å)	12
RNA stems	16	Weak	(0.0–3.5 Å)	13
		3D strong	(0.0–5.0 Å)	57
		3D weak	(0.0–7.0 Å)	41
Total NOE constraints	**123**			

rmsd profiles for the 50 final structures are shown in Figure 5. Inspection of these plots allows straightforward identification of an ensemble of converged structures (in this case structures 1–31), corresponding to clear plateaux in the profiles; structures beyond structure 31 provide a significantly poorer fit to the experimental data. All of the superpositions, analysis and statistics that follow refer to the ensemble of structures 1–31.

Intermolecular hydrogen bonds were identified by analysing the converged structures using the program 'hbplus' (McDonald and Thornton, 1994), accepting geometries where the donor–acceptor (D–A) distance was less than 4 Å provided also that the proton–acceptor (H–A) distance was shorter than the D–A distance. Hydrogen-bonding interactions were considered to be present when at least 50% of the converged structures met these criteria. Hydrophobic interactions were identified by searching for all intermolecular carbon-carbon distances of less than 4 Å. This cutoff is just above the longest carbon–carbon distance for two CH groups in van der Waals contact (the C–C distance is 3.8 Å when all four atoms are arranged linearly, otherwise it is shorter). Hydrophobic interactions were considered to be present when at least 65% of the converged structures met this criterion. Electrostatic interactions are more difficult to define by NMR; those discussed in the text were proposed based on short (<5 Å) distances between Nζ of Lys (or Nη of Arg) and a non-bridging phosphate oxygen in at least 13% of the converged structures. Note that electrostatic components of the X-PLOR (Brünger, 1990) force

field were inactive throughout the structure calculation protocol, to avoid any bias in the identification of intermolecular electrostatic interactions and salt bridges.

Results and Discussion

Assignments of the U1A–RNA complex

Assignments of the protein resonances in the complex are essentially complete and are available as supplementary material. All backbone amide groups were assigned, but some side-chain signals remain unassigned (Lys^{20} and Lys^{28} Hϵ and Cϵ; Lys^{50} Hδ, Cδ, Hϵ and Cϵ; Gln^{54} Hδ; Phe^{56} Cϵ and Cζ; Phe^{59} Hζ, Cϵ and Cζ; Met^{72} Hγ and Cγ; Phe75 Hϵ, Cϵ, Hζ and Cζ; Pro^{81} Hγ and Cγ; Phe^{101} Cϵ; no stereoassignments were made, and in a number of methylene groups only one proton was assigned). All the arginine Nϵ and Hϵ signals (except from Arg^{83}) were assigned using a 2D-HNCACB experiment; the chemical shift of Arg^{52} Hϵ (5.94 ppm) is unusual, and probably reflects proximity to the aromatic ring of an RNA base. The Hζ signal of Lys^{80} was also assigned.

Some resonances were difficult to assign because they had unusual shifts and/or broad line widths. These include Pro8 Hα at 1.66 ppm, Ile40 Hγ at 0.01 ppm and Ser^{48} Hβ at 2.30 ppm. The entire Ser^{91} spin system was broadened and difficult to assign, as were the aromatic rings of Phe^{56} and Tyr^{13}, probably because intermolecular stacking causes them to flip at an intermediate rate on the chemical shift time scale (Figure 6). Three hydroxyl protons were also assigned (Tyr^{13} OH at 10.38, Ser^{46} OH at 6.80 and Ser^{48} OH at 6.84 ppm); the slow rate of solvent exchange of these protons suggests that they are hydrogen-bonded.

Comparison of protein chemical shift values in the free (U1A-117) (Avis et al., 1996) and the bound state (U1A-102) provides a qualitative indication of the regions of the protein that interact with RNA. Most chemical shift changes in the protein backbone (Figure 7a) occur in the four strands of the β-sheet and in loop 3 (residues 45–53, between strands β2 and β3), in the loop between strand β4 and helix C (residues 86–92). Changes observed in the C-terminus of helix A (residues 29–33) are probably due to the Tyr^{31}His mutation. The absence of any clear backbone chemical shift differences in either loop 1 (connecting strand β1 and helix A) or in the C-terminal helix contrasts to the large chemical shift changes observed in loop 3 and the β4–helix C loop. The largest side-chain chemical shift differences are observed for the hydrophilic residues Asn^{9}, Asn^{15}, Ser^{46}, Ser^{48}, Gln^{85} and Lys^{96} and the hydrophobic residues His^{10}, Thr^{11}, Tyr^{13}, Leu^{44}, Leu^{49}, Phe^{56}, Ile^{58}, Val^{62}, Ala^{87}, Ile^{93} and Ile^{94} (Figure 7b). The free U1A-117 assignments were used in this comparison because several amide group resonances are missing in the region of loop 3 for the free U1A-102 spectra.

The assignments of the RNA in the complex were reported previously (Gubser and Varani, 1996). They are essentially complete, except the H5′/H5″/C5′ signals of A22, G25, A39 and U41–C45, and the amino resonances of part of the single-stranded region (A39–A44). U40, U41 and G42 imino signals were observed and assigned in the complex but not in the free RNA, suggesting that they are probably hydrogen-bonded in the complex. Most chemical shifts remain unchanged for the two helical stems of the RNA upon complex formation, excepting the G25•C38 and G23•C46 base pairs and some resonances in stem 2 (C21 and A22). Very large chemical shift changes were observed throughout the single-stranded loop (A39–C45) and in A24.

The chemical shift changes are entirely consistent with those reported for studies of RNP-protein complexes with RNA (Görlach et al., 1992; Howe et al., 1994; Kanaar et al., 1995). In each case, chemical shift changes occur in the four-stranded β-sheet and the C-terminus of the protein, while α-helices A and B are not affected. Chemical shift changes for loop 3 could not be assessed in previous studies because assignments were unavailable for that region (Görlach et al., 1992; Howe et al., 1994; Kanaar et al., 1995). Only the large chemical shift changes observed for the N-terminal region of the RNP domain of hnRNPC (Görlach et al., 1992) differ from the present study, suggesting perhaps a different mode of interaction of hnRNPC with its RNA target.

Composition of the final constraint list

The structure determination of the complex depended to a large extent on identification of the 123 intermolecular interproton distance constraints listed in Table 1. These constraints are not evenly distributed either amongst the nucleotides or the amino acids that comprise the intermolecular interface. Nucleotides G25, U40, G42, C43, A44 and C45 are each constrained by 10–30 intermolecular NOE contacts; A24, A39 and U40 each by 5–10 constraints, and C46 by a single constraint; no intermolecular constraints were unambiguously identified for either G23 or C38 (Figure 8).

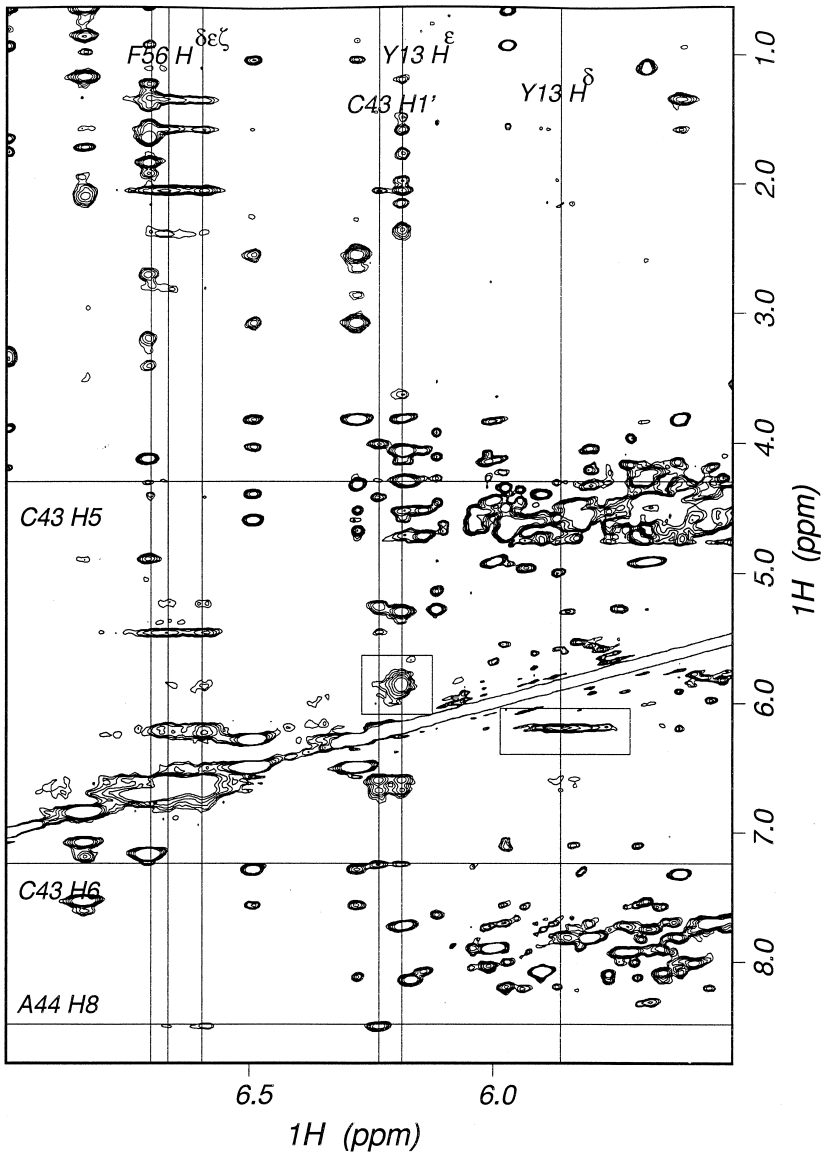


Figure 6. Section of a 2D-[^{1}H-^{1}H]-NOESY spectrum of the complex in D_2O, showing resonances from the aromatic rings of Tyr^{13} (δ and ϵ) and Phe^{56} (δ, ϵ and ζ). Several of these resonances are broadened, most probably due to a slow ring flip rate.

In the protein, only 18 residues display intermolecular NOE contacts: Thr^{11}, Tyr^{13}, Asn^{15} and Asn^{16} in strand β1, Leu^{44} in β2, Phe^{56} in β3 and Lys^{80} and Gln^{85} in β4, Ser^{48}-Gly^{53} in loop 3 and Lys^{88}, Thr^{89}, Ser^{91} and Asp^{92} in the loop between β4 and helix C.

Only 25% (31 out of 123) of the intermolecular NOE constraints involve exchangeable protons. These include six protein main-chain amides (Asn^{16}, Ser^{48}, Leu^{49}, Lys^{50}, Gly^{53} and Asp^{92}), two side-chain amides (Asn^{15} and Gln^{85}), Tyr^{13} OH, Arg^{52} Hϵ and Lys^{80} NζH_3, and the U40, U41, G42 imino and C45 RNA amino protons. This proportion is small, and would undoubtedly have been higher if exchangeable signals from the ribonucleotide exocyclic aminos and 2′-hydroxyls, and from lysine and arginine side chains could have been assigned. Such NOE contacts with exchangeable protons are amongst the most important but also the most difficult to obtain. For example, the side-chain amide signals of Gln^{54} could not be assigned so this residue was not precisely positioned in the structure; in contrast, U40 and U41 were precisely defined only after their imino protons had been de-

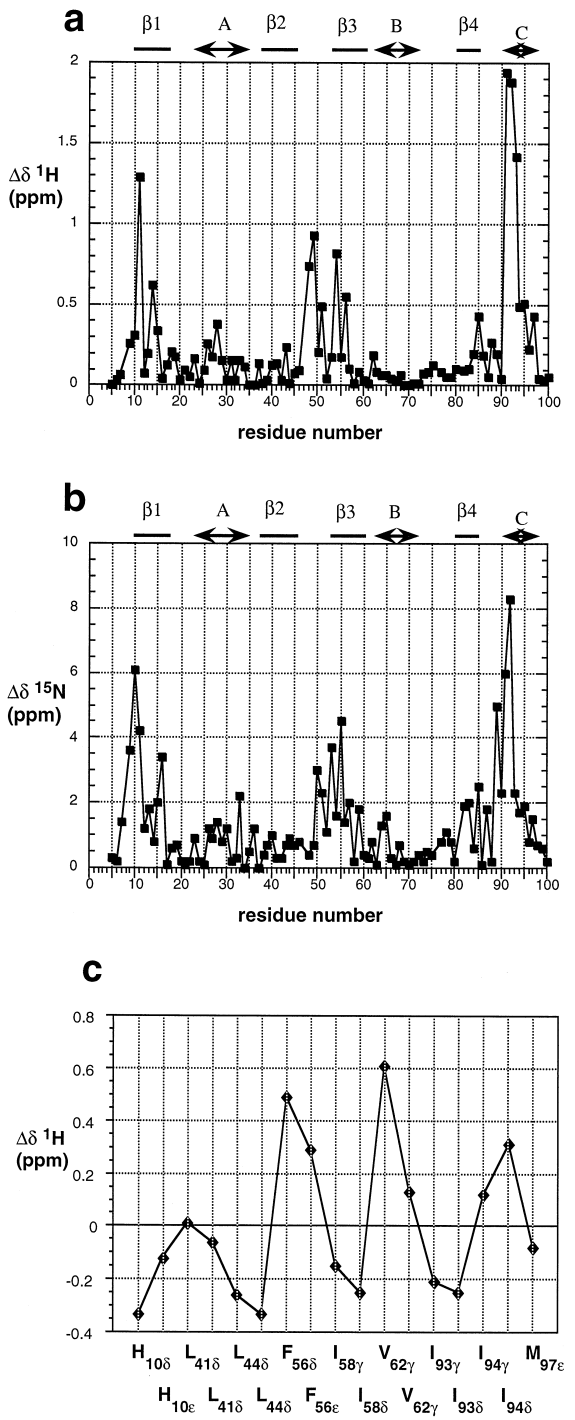


Figure 7. Chemical shift changes between the free (A-117) and the bound (A-102) states of the protein. (a) ^{1}H of backbone amide groups, (b) ^{15}N of backbone amide groups and (c) hydrophobic side chains of residues involved in positioning helix C.

tected and assigned (for which purpose a new labelled RNA sample was specifically prepared).

For the protein, the 1710 intramolecular NOE-derived distance constraints (≈17 per amino acid) and 42 constraints for 21 hydrogen bonds represent 67% of the total of distance constraints in the complex. For the RNA, the 591 intramolecular NOE-derived distance constraints (≈20 per nucleotide), 110 dihedral angles and 25 hydrogen-bonding constraints (in the double-helical stems only) represent 28% of all constraints in the complex. The 123 intermolecular NOE contacts represent only a small fraction (≈5%) of the 2602 NMR-derived constraints that were used for the calculation of the whole complex (Table 2), but of course they play a disproportionately important role in defining the structure.

Most intramolecular NOE correlations found in the bound protein and RNA components were already present and assigned in the free protein. However, some genuine structural differences are revealed by the presence of several long-range connectivities that are unique to either the complex (≈40 NOE contacts) or the free protein (≈80 NOE contacts). In the complex, several NOE cross peaks are observed between His^{10}, Leu^{41}, Ile^{58}, Val^{62} and Ile^{93}, Ile^{94} and Met^{97}; these are not observed in the free protein, instead there are NOE cross peaks between residues 93–97 and Tyr^{13}, Leu^{44} and Phe^{56}. These NOE cross peaks position helix C in very different orientations in the free and bound states of the protein (Allain et al., 1996; Avis et al., 1996). Additional NOE contacts between Asn^{16} and Pro^{81}, Leu^{17} and Leu^{26}, Glu^{19} and Ser^{46}, and between Ser^{46} and Met^{51} are observed in the free protein (A-117) but not in the complex. Finally, Asn^{9}, Thr^{11} and Ile^{12} have slightly different intramolecular NOE connectivities in the free and bound proteins. For the RNA, 32 of 65 constraints that differ between the free and the bound forms reflect genuine structural differences. Of these, 16 constraints involve the C43 nucleotide. C43 had very broad resonances in the free RNA, displayed only two NOE contacts (Gubser and Varani, 1996) and was the least well defined nucleotide in the free RNA structure. In contrast, C43 is involved in 29 intermolecular NOE contacts in the complex, in addition to 16 intramolecular NOE contacts, making C43 the most highly constrained nucleotide at the RNA–protein interface. Other intramolecular distance constraints observed only in the complex involve the imino protons of G23, U40 and G42 (10 constraints); these resonances are in slower exchange with solvent in the complex than in the free RNA. The NOE con-

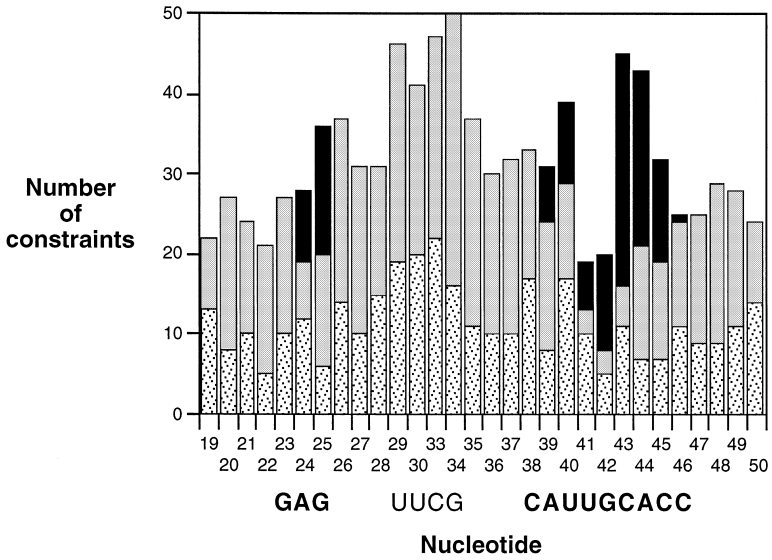


Figure 8. Plot showing the distribution of different types of NOE constraints for the RNA in the complex. Intra-nucleotide NOE constraints are shown in 'speckle', inter-nucleotide NOE constraints are shown in grey and intermolecular NOE constraints to the protein are shown in black.

Table 3. Structural statistics for the final ensemble of 31 structures

(a) Structural statistics		
NOE violations		
Number $>$ 0.2 Å	4 ± 1	
Maximum violation	$<$ 0.5 Å	
Angle violations		
Number $>5°$	1.6 ± 1	
Mean deviation from ideal covalent geometry		
Bond length	0.05 Å	
Bond angles	0.80°	
Impropers	0.45°	
(b) Mean rms deviations from average structure (Å)		
Protein		
All ordered residues (7–98)		
Backbone		0.54 ± 0.07
Heavy atoms		0.98 ± 0.11
RNA (all heavy atoms)		
All ordered regions		2.11 ± 0.72
Upper stem (G25–C38)		1.03 ± 0.37
Lower stem (G20–G23, C46–C49)		0.80 ± 0.11
Protein binding site (G23–G25, C38–C46)		1.01 ± 0.16
Single-stranded loop (A39–C45)		0.87 ± 0.11
Protein–RNA (all heavy atoms)		
All ordered regions (7–98, G20–C49)		2.03 ± 0.61
Complex interface		1.08 ± 0.19

straint between G42 H1 and U40 H3 positions G42 and U40 in close proximity. The remaining six constraints observed only in the complex are important for positioning A24, A44, C45 and C46.

Determination of the structure of the protein–RNA complex

The restrained molecular dynamics protocol used previously for the free protein (Nilges et al., 1988; Avis et al., 1996) and that used for the free RNA (Wimberly, 1992; Gubser and Varani, 1996; Varani et al., 1996) were both unsuccessful when applied directly to the complex. When the 'protein' protocol was used, the initial 'search phase' of the protocol (i.e. the first 10 000 dynamics steps at 1000 K) successfully located the global fold of the complex, but during the second step of the protocol (progressive switching from 'soft' to 'square' potential for NOE interactions, progressive increase of the van der Waals radii and a decrease of the step size for the molecular dynamics from 7 to 5 fs) the various energy terms and the temperature rose in an uncontrolled fashion. Clearly, one or more of the changes occurring during this phase of the protein protocol is too extreme or too rapid for the calculation of the RNA component. When the 'RNA' protocol was used alone, the calculation failed almost immediately, perhaps because it was swamped by relatively large energy terms originating from the randomized protein component. However, when the search phase from the protein protocol was followed by the entire RNA protocol, the calculation succeeded. This combined protocol was used to derive the structures of the complex presented here. It accepts a starting structure comprising both components with randomized backbone angles as input, and acts simultaneously on both components to produce the complete structure of the complex in one operation; this has the advantage that it uses no docking step which could introduce bias or reduce sampling of conformational space. No attempt was made to optimize details of this 'combined' protocol, so it may be unnecessarily long as presently implemented (it requires 4 h per structure on a DEC 2100 4/275 Alpha-server).

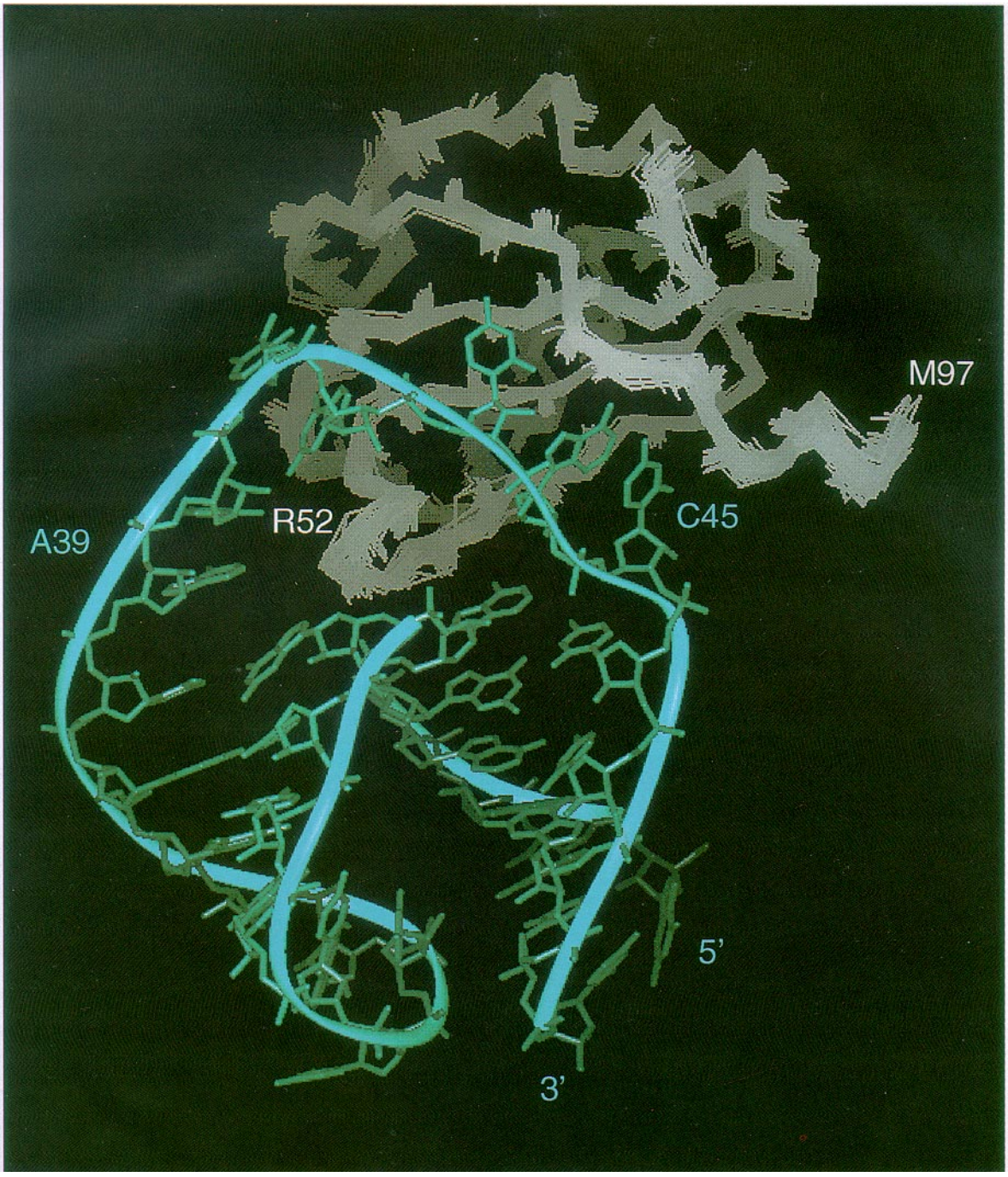
Each of 50 starting structures (each comprising both protein and RNA components) entering the protocol led to a final structure of the complex. The total energy (E_{total}) of the final structures varied from 600 to 10 000 kcal/mol and the NOE energy (E_{NOE}) from 50 to 4000 kcal/mol (Figure 5a). Of the 50 final structures, 31 have comparable E_{total} values ($\approx$600 kcal/mol) and E_{NOE} values ($\approx$50–90 kcal/mol); these 31 structures define the ensemble of converged structures used for reporting the structural statistics, as judged by inspection of the energy-ordered profiles and rmsd profiles (Figure 5). This convergence rate (62%) is lower than that for the free protein (86%) (Avis et al., 1996), but is comparable to those for other RNA structures determined in this laboratory (Allain and Varani, 1995; Gubser and Varani, 1996; Varani et al., 1996).

The quality of the structure can be judged by the low number of constraint violations, the low deviations from ideal covalent geometry and the good precision (Figure 9). Rmsd values for the ensemble of 31 converged structures are reported in Table 3. The precision of the protein backbone (0.54 Å) and heavy atoms (0.98 Å), of the RNA seven-nucleotide loop (0.87 Å) and of the entire RNA–protein interface (1.08 Å) allows a detailed analysis of intermolecular interactions.

Structure of the U1A–RNA complex

As described in the initial report of the structure of this complex (Allain et al., 1996), the polyadenylation inhibitory element (PIE) RNA interacts with the surface of the four-stranded β-sheet and with three protruding loops (β1–helix A, β2–β3 and β4–helix C) of the U1A protein. The RNA is severely kinked at the internal loop site, and the single-stranded nucleotides are splayed out across the surface of the β-sheet. Loop 3 of the protein, which connects strands β2 and β3 of the β-sheet, protrudes through the hole in the RNA defined by the single-stranded nucleotides (A39–C45 and A24) and the two base pairs at the start of the two stems (G25•C38 and G25•C46). All of the unpaired nucleotides (A39 to C45 and A24) are involved in intra- and/or intermolecular stacking interactions.

Nucleotides G42, C43, A44 and C45 stack with Gln^{54}, Tyr^{13}, Phe^{56} and Asp^{92}, respectively (Figure 9c), while the negatively charged phosphates of the RNA backbone are directed away from the protein into solution. These intermolecular stacking interactions have counterparts in tRNA-synthetases and other RNA–protein complexes, and appear to be a common feature of RNA recognition (Rould et al., 1989, 1991; Caverelli et al., 1993; Belrhali et al., 1994; Valegárd et al., 1994). The structure of the complex is strikingly different from an early model of the related hairpin complex based on biochemical data (Jessen et al., 1991), which predicted that the protein–RNA interaction would be predominantly mediated by the



a

Figure 9. Superpositions of the 31 converged structures of the complex of U1A protein and the PIE-RNA, determined by NMR. Panel (a) shows a protein backbone superposition (in grey; superposed on average coordinates for the whole protein backbone), together with a single RNA structure (cyan; the RNA shown is part of the lowest energy complex structure); panel (b) shows an RNA backbone superposition (in cyan; superposed on the average RNA structure for all heavy atoms in the interfacial region), together with a single protein structure (white; the protein backbone shown is again part of the lowest energy complex structure); and panel (c) shows the interfacial region, superposed as in (b) (the RNA is coloured cyan with phosphate groups in dark blue, while the protein is coloured yellow for interfacial backbone atoms, white for interfacial hydrophobic residues, green for interfacial hydrophilic residues and red for the backbone ribbon).

b

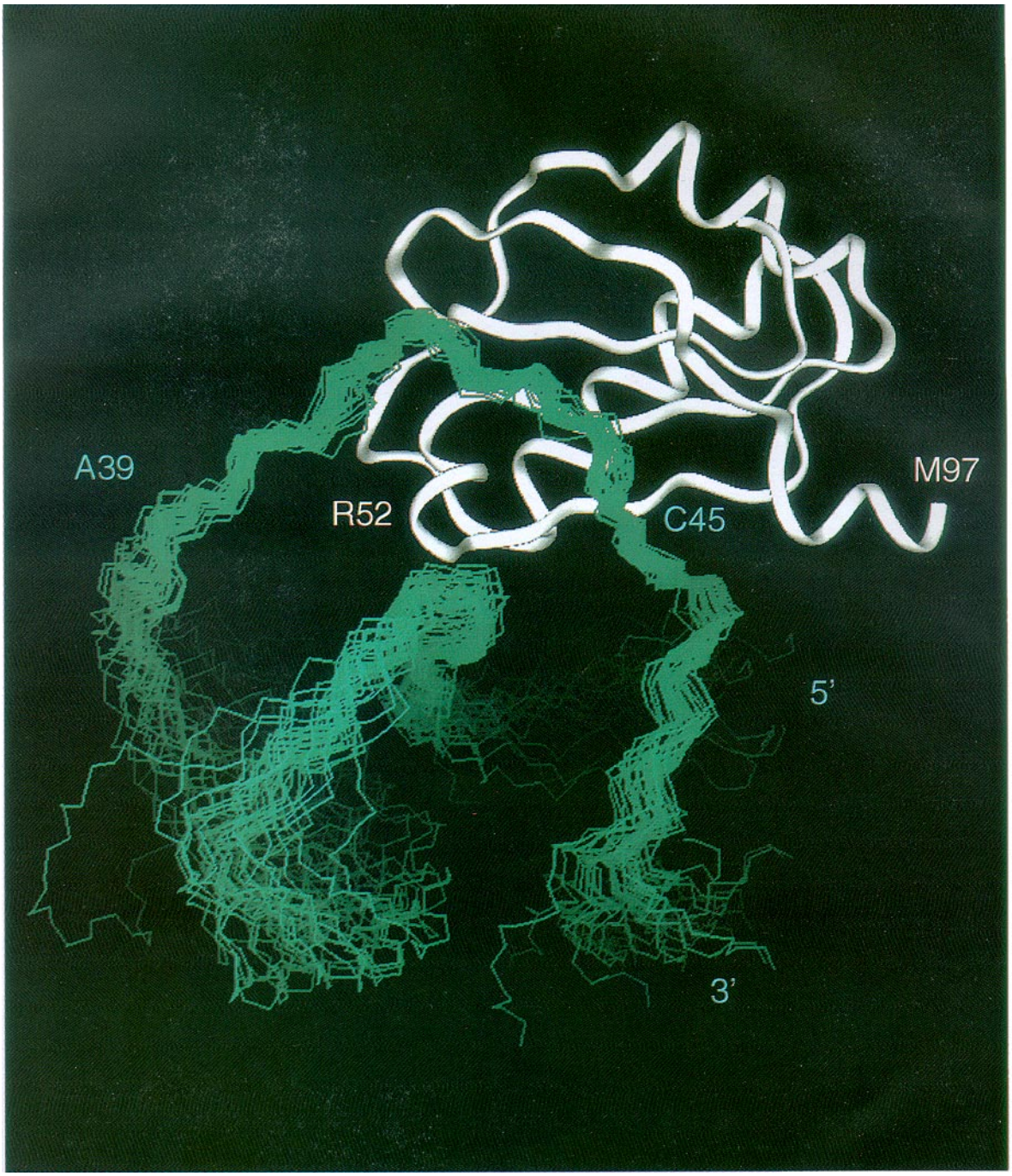


Figure 9. Continued.

c

Figure 9. Continued.

phosphate backbone with the RNA bases exposed to solvent. In contrast, the architecture of the actual structure of the complex resembles somewhat an enlarged protein structure, with a 'core' made up of protein hydrophobic side chains and RNA bases in the interior, while positively charged protein side chains and negatively charged RNA backbone phosphates are on the surface (Figure 9c).

Intermolecular RNA–protein interactions can be divided into four general classes: stacking, hydrophobic, hydrogen-bond and electrostatic interactions. By hydrophobic interactions here we mean close contacts between hydrophobic amino acid side chains of the protein and CH groups of the RNA bases and sugars (although we do not mean to imply such interactions are truly hydrophobic in a strict thermodynamic sense). In each case, a statistical analysis of all 31 converged structures was performed to identify intermolecular interactions. Since NMR detects short interproton distances (<5–6 Å), stacking, hydrophobic and (to a lesser extent) hydrogen-bonding contacts can be defined more precisely than electrostatic interactions or salt bridges. Tables 4a–c list all observed intermolecular hydrophobic (including stacking), hydrogen-bond and electrostatic contacts in the NMR structures.

Most intermolecular contacts reported in Table 4 have been described previously (Oubridge et al., 1994; Allain et al., 1996). However, a number of novel interactions are only identified in this refined structure. Most of the newly identified contacts involve loop 3 of U1A and A24 base and the sugar ring of G25. These contacts include three hydrogen bonds (Val^{45} carbonyl and Arg^{47} and Leu^{49} amides), seven hydrophobic contacts (from Ser^{46} and Ser^{48} side chains) and one electrostatic interaction (from the Arg^{47} side chain). Other interactions between loop 3 and the RNA include contacts involving Ser^{48} hydroxyl, Leu^{49}, Lys^{50}, Met^{51} and Arg^{52} side chains, and have already been described (Allain et al., 1996). Overall, residues within loop 3 form 17 hydrophobic contacts, nine hydrogen bonds and two salt bridges (Table 4), reinforcing the importance of this region of the RNP domain in RNA recognition.

Another set of interactions not previously reported connect side chains of basic residues on the protein to phosphates in stem 2 of the RNA (Figure 1); Lys^{23}, Lys^{96} and Arg^{47} contact respectively the phosphates of A22, C46 and G23 (Table 4c). This helical region of the RNA has no counterpart in stem-loop II of U1 snRNA (Oubridge et al., 1994; Nagai et al., 1995).

Table 4a. Statistics for intermolecular distances between hydrophobic groups of the protein and CH groups of the RNA in the final ensemble of 31 structures

RNA	Protein	%	C–C distance (Å)
A24 C6	S46 Cα	94	3.65 ± 0.23
A24 C6	S46 Cβ	90	3.67 ± 0.39
A24 C5	S46 Cα	90	4.21 ± 0.16[a]
A24 C5	S46 Cβ	71	3.85 ± 0.25
A24 C2′	S48 Cβ	100	3.86 ± 0.12
A24 C3′	S48 Cβ	100	3.72 ± 0.16
G25 C5′	S48 Cβ	94	3.91 ± 0.23
G25 C8	L49 Cα	100	3.83 ± 0.11
G25 C8	L49 Cβ	100	3.46 ± 0.09
G25 C4′	L49 Cβ	100	4.11 ± 0.11[a]
G25 C1′	L49 Cβ	100	3.57 ± 0.15
G25 C1′	L49 Cδ1	100	3.75 ± 0.09
A39 C2	L49 Cδ1	100	3.79 ± 0.21
U40 C1′	L49 Cδ2	90	3.80 ± 0.21
G42 C2	R52 Cβ	94	3.60 ± 0.22
C43 C6	K88 Cϵ	100	3.75 ± 0.33
C43 C5	K88 Cϵ	74	3.64 ± 0.18
C43 C4	K88 Cγ	65	3.85 ± 0.26
C43 C2	K88 Cγ	87	3.55 ± 0.28
C43 C2′	K88 Cϵ	100	3.82 ± 0.25
A44 C1′	M51 Cϵ	87	3.74 ± 0.30
A44 C1′	F56 Cϵ	100	3.85 ± 0.25
A44 C1′	F56 Cζ	100	3.73 ± 0.15
A44 C4′	M51 Cϵ	81	3.93 ± 0.30
A44 C2	L44 Cδ1/2	100	4.18 ± 0.11[a]
C45 C4	S90 Cα	100	3.85 ± 0.16
C45 C1′	L44 Cδ2	100	3.87 ± 0.27
G42	Q54	87	Stacking
C43	Y13	100	Stacking
A44	F56	100	Stacking
C45	D92	80	Stacking

The cut-off C–C distance for acceptance of a given hydrophobic contact was 4 Å, except for the interactions marked [a] where it was set to 4.5 Å. The interactions marked 'stacking' were assessed visually in the structures, and are included here for completeness.

The unusual chemical shifts of resonances such as Arg^{52} Hϵ, Ser^{48} Hβ and Asp^{92} NH are due to interactions with the RNA. For example, Arg^{52} interacts with the G25•C38 base pair and with A39, so the upfield shift of this resonance probably results from ring current effects from these RNA bases. Intermolecular stacking on C43 and A44 could be responsible for the slow flip rate of the Tyr^{13} and Phe^{56} aromatic rings. The large chemical shift differences observed between the bound and free protein in loop 3 and in the β4–helix C loop (Figure 7) are consistent with the

Table 4b. Statistics for intermolecular hydrogen bonds in the final ensemble of 31 structures

RNA	Protein	H–A distance (Å)	D–A distance (Å)	D–H–C angle (°)	%	H-bond in stem-loop H/U1A complex (X-ray)
A24 N6-H	Val^{45}O	2.6 ± 0.4	3.2 ± 0.5	121 ± 16	90	na
A24 N7	Arg^{47} N-H	2.7 ± 0.1	3.6 ± 0.2	158 ± 12	84	na
G25 OP	Ser^{48} Oγ-H	*	3.1 ± 0.4	*	87	na
G25 O4′	Leu^{49} N-H	2.6 ± 0.2	3.4 ± 0.2	141 ± 4	100	na
G25 N7	Arg^{52} Nη-H	2.9 ± 0.4	3.7 ± 0.3	135 ± 24	55	Same
A39 N1	Arg^{52} Nη-H	2.7 ± 0.4	3.3 ± 0.3	124 ± 23	77	Same
U40 N3-H	Glu^{19} Oϵ1/2	2.7 ± 0.4	3.2 ± 0.3	113 ± 21	87	Same
U40 O2 G42 N2-H		2.2 ± 0.5	2.9 ± 0.2	122 ± 16	100	Same
U41 N3-H	Asn^{16} Oδ	2.8 ± 0.3	3.6 ± 0.3	134 ± 20	61	Same
G42 N1-H	Arg^{52}O	2.8 ± 0.4	3.0 ± 0.3	88 ± 24	97	G42 N1-H to Glu^{19} Oϵ1/2
G42 N2-H	Leu^{4}9 O	2.5 ± 0.1	3.1 ± 0.4	118 ± 14	97	G42 N2-H to Glu^{19} Oϵ1/2
G42 N7	Asn^{15} Nδ2-H	2.6 ± 0.2	3.4 ± 0.2	139 ± 9	58	Same
G42 O2′-H	Lys^{50}O	*	3.4 ± 0.5	*	87	Same
C43 N3	Lys^{88} N-H	3.0 ± 0.4	3.5 ± 0.2	140 ± 10	90	Same
C43 N4-H	Tyr^{86} O	2.3 ± 0.3	3.3 ± 0.3	155 ± 12	87	Same
C43 O2	Lys^{88} N-H	2.5 ± 0.1	3.4 ± 0.2	154 ± 14	87	Thr^{89} N-H to C43 O2 via H_2O
A44 N6-H	Thr^{89} O	3.4 ± 0.3	3.9 ± 0.2	113 ± 11	94	Same
C45 N4-H	Asp^{90} O	2.1 ± 0.3	2.6 ± 0.2	109 ± 13	100	Same
C45 O2	Ser^{91} Oγ-H	*	3.5 ± 0.4	*	51	C45 O2 to Asp^{92} O via H_2O
C45 N3	Asp^{92} N–H	2.5 ± 0.1	3.3 ± 0.1	142 ± 14	100	Same

One intramolecular hydrogen bond in the RNA (U40 O2 – G42 N2-H) is also included, as this interaction was only detected for the complex and was found during the same analysis. The '%' column indicates the percentage of structures for which the indicated hydrogen bond was accepted as being present (see text for cut-off criteria). na in the 'X-ray' column indicates not applicable; these are cases where the stem-loop II RNA has no counterpart of the interfacial nucleotide that forms a particular intermolecular hydrogen bond in the PIE-RNA complex. No statistics are given for distances or angles directly involving hydrogens that are linked to the rest of the structure via an adjacent rotatable bond and where no corresponding ^{1}H resonance was observed (e.g. Ser OH and 2′ OH of RNA sugars; marked * in the table); positions of these atoms in the calculated NMR structures are effectively randomized by rotations of the adjacent rotatable bond (Cβ-Oγ for Ser, C2′-O2′ for RNA sugars). In addition, there are nine hydrogen bonds in the stem-loop II X-ray structure that are rare (<35% occurrence; data not shown) in the NMR structure; these comprise Arg^{52} NηH1/2 to G25 O6, Lys^{80} NηH^3 to U41 O4, Asn^{16} NH to G42 O6, G42 N1-H to Glu^{19} Oϵ1/2, G42 amino to Glu^{19} Oϵ1/2, C43 amino to Gln^{85} Oϵ1, Lys^{88} NζH_3 to C43 O2′, Lys^{50} NζH_3 to A44 OP and Ser^{91} Oγ-H to A44 N1.

involvement of these regions of the protein in many intermolecular hydrogen bonds. The RNA chemical shift changes perfectly map the RNA binding site, but chemical shift changes do not necessarily imply direct intermolecular interactions. A number of chemical shift changes can be attributed to protein conformational changes, particularly the movement of helix C (Asn^9, His^{10}, Leu^{41}, Ile^{58}, Val^{62}, Ile^{93}, Ile^{94} and Met^{97} side chains). The unusually slow rate of exchange of resonances such as Tyr^{13} and Ser^{48} hydroxyls, Lys^{80} amine and U40, U41 and G42 imino or C45 amino in the bound RNA can be explained by the formation of hydrogen bonds in the complex; however, not all protons identified as hydrogen-bonded based on their positions in the calculated structures could be observed in the spectra.

Comparison with the crystal structure of the U1A–hairpin RNA complex

The expected similarities between the present structure and the 1.92 Å crystallographic structure of the related U1A–hairpin complex are confirmed by the comparison of the two structures. As originally suggested, the single-stranded nucleotide sequence 5′AUUGCAC3′ is recognized in a nearly identical manner in the two complexes, despite the difference in secondary structural context (Oubridge et al., 1994; Nagai et al., 1995). The rmsd between the average structure of the ensemble of NMR structures and the

Table 4c. Statistics for distances corresponding to intermolecular electrostatic interactions in the final ensemble of 31 structures

RNA phosphate O1 or O2	Protein Lys Nϵ or Arg Nη	% <5 Å
A22	Lys^{23}	54
G23	Lys^{23}	26
G23	Arg^{47}	32
C43	Lys^{88}	35
A44	Lys^{50}	23
C46	Lys^{96}	13

crystal structure is 1.13 Å for these seven nucleotides, 1.29 Å for the portion of the protein–RNA interface involving these nucleotides and 1.02 Å for the protein backbone. These values demonstrate that NMR can determine the RNA–protein interface not only with high precision but also to an accuracy of ≈ 1.3 Å.

However, a more detailed inspection reveals that details of the array of intermolecular hydrogen bonds are somewhat different. Among the intermolecular hydrogen bonds involving the seven single-stranded nucleotides (A39–C45), 12 are common to both structures, four are only present in the NMR complex and nine are present in the crystal structure but rare (<35% occurrence; data not shown) in the NMR structure. In addition, the crystal structure contains five bound water molecules at the protein–RNA interface mediating intermolecular contacts. Examination of the four hydrogen bonds found only in the NMR structure reveals that several of the donors and acceptors are groups that interact with interfacial water molecules in the crystal structure. This indicates that the four apparent 'NMR only' hydrogen bonds were probably found in the analysis only because the acceptance criteria used to identify hydrogen bonds in the ensemble of NMR structures were sufficiently loose that they could be satisfied by water-mediated interactions as well as by direct hydrogen bonds. Future NMR experiments for the observation of interfacial water molecules may clarify this issue.

Comparison of the two structures in the region of loop 3 reveals further differences. In the PIE-RNA complex, protein residues Val^{45} to Leu^{49} form four hydrogen bonds and extensive hydrophobic contacts with G25 and the unpaired A24 nucleotide on the RNA (Table 4). The positions of these nucleotides correspond approximately to those of nucleotides G16 and C15 in the stem-loop II complex. Detailed comparison in this region of the structures shows the following: (i) In the PIE-RNA complex, the ring of A24 points towards the protein backbone and is involved in both intermolecular hydrogen bonds and hydrophobic interactions (Table 4), whereas in the stem-loop II complex the corresponding ring of C15 is directed away from the protein and shows no intermolecular interactions. This may reflect poor ordering and crystal packing interactions in the U13–C15 region in the X-ray structure (Oubridge et al., 1994; Nagai et al., 1995). (ii) In the PIE-RNA complex, G25 O4′ is hydrogen-bonded to Leu^{49} NH and the G25 non-bridging phosphate oxygen hydrogen-bonds to Ser^{48} OH (Table 4), whereas in the stem-loop II complex G25 O4′ is not hydrogen-bonded, the G25 non-bridging phosphate oxygen hydrogen-bonds to Leu^{49} NH and Ser^{48} OH forms an intramolecular hydrogen bond to Ser^{46} OH. (iii) In the PIE-RNA complex, Val^{45} and Lys^{23} hydrogen-bond with A24 and A22, respectively, while Ser^{46} contacts A24 (Table 4), whereas in the stem-loop II complex these protein residues form only intramolecular interactions. These distinctions are directly supported by observed differences between NMR spectra of the two complexes. In particular, the backbone amide groups of loop 3 (Ser^{46}–Arg^{52}) give clear cross peaks only in ^{1}H-^{15}N HSQC spectra of the complex with PIE-RNA; the corresponding cross peaks in the spectrum of the stem-loop II complex are all significantly broadened.

Overall, these comparisons show that U1A–RNA interactions in the PIE-RNA complex are not limited to the seven conserved nucleotides (5′AUUGCAC3′) and the G•C base pair closing the stem, as was described for the case of the stem-loop II complex (Oubridge et al., 1994; Nagai et al., 1995). The PIE-RNA complex shows interactions from the protein to A24, A22 and C46, and the interactions to G25 (whose counterpart in the stem-loop II RNA is G16) are also very different. Although the protein backbone of loop 3 maintains the same helical shape in the free protein and in the two complexes, its side chains (Ser^{46}, Ser^{48}, Arg^{47}, Lys^{23}) interact with the RNA differently in the two complexes. These observations have important implications for understanding the molecular origin of RNP-RNA specificity, as discussed elsewhere (Allain et al., 1997).

Comparison with recent NMR structures of peptide–RNA complexes

In addition to the present structure, four structures of peptide–RNA complexes from immunodeficiency

viruses have been recently determined by NMR (Puglisi et al., 1995; Ye et al., 1995, 1996; Battiste et al., 1996). In each case, the RNA molecule was of comparable size to the 3′UTR RNA (30–35 nucleotides), but much shorter, unstructured peptides of 14–22 amino acids were studied. The RNAs were labelled with ^{15}N and ^{13}C in all cases, but the peptide was only labelled in one case (Battiste et al., 1996).

The quality of the spectra of the peptide complexes allowed nearly complete spectral assignments of the isotopically labelled RNAs and peptides in all cases, but the number of NMR-derived distance constraints varied considerably. The average number of intramolecular RNA interproton distance constraints for the different complexes varied between 11 and 25 constraints per nucleotide; these numbers are comparable to the present complex (19 constraints per nucleotide). The average number of intramolecular peptide interproton distance constraints for the same four complexes was between four and 10 constraints per amino acid, much lower than the 19 constraints per amino acid obtained for the U1A protein in the present complex. In the peptide–RNA complexes the average number of intermolecular distance constraints was between two and 10 per interfacial residue while an average of nine intermolecular distance constraints per interfacial residue was collected for the U1A–RNA complex (calculated as $2\times$ (number of interfacial constraints) $\div$ (number of interfacial amino acid residues $+$ number of interfacial nucleotides)). Thus, the average number of distance constraints obtained per residue in the U1A complex is at least as large as those reported for peptide–RNA complexes, despite the much higher molecular weight.

U1A uses the surface of the β-sheet to bind an RNA single-stranded region, whereas arginine-rich peptides are inserted in the open major groove of a double-helical RNA. The extensive surface complementarity is a striking feature common to both types of structure. In the present complex, surface complementarity is achieved through the folding of the single-stranded nucleotides against the surface of the β-sheet and through insertion of the protein loop 3 in the hole formed by the RNA backbone. In the RNA–peptide complexes, this is achieved instead through the folding and deep penetration of the arginine-rich peptides into the RNA major groove. The resulting co-penetration of nucleotides and amino acids explains the equally high density of intermolecular constraints found in the U1A complex and the peptide–RNA complexes.

Conclusions

A number of technical problems had to be overcome in order to determine the high-resolution structure of the 22 kDa complex between the RNA-binding domain of human U1A protein and part of the polyadenylation inhibition element from U1A mRNA. Sample solubility and long-term stability were improved by modifying the protein purification to remove contaminating nuclease activity, by site-directed mutagenesis and by optimization of solvent conditions and sample handling. Spectral assignments of the RNA and protein components were greatly facilitated by the availability of complete assignment sets for the unbound counterparts (Avis et al., 1996; Gubser and Varani, 1996). Because of the high molecular weight and relatively fast relaxation of the present complex, assignments relied heavily on NOE distance information. The NOE-based distance constraint set was constructed using the constraint lists for free RNA and protein components as a template. Thus, only a short time (approximately 1–2 months) was needed to obtain the first, low-quality structure of the complex after completion of the free RNA and protein structures and assignment of the resonances of the complex. Preliminary structures were used in an iterative fashion to identify the critical intermolecular NOE constraints that define the geometry of the interface and these extra constraints resulted in a significant improvement in the precision for the protein–RNA interface. A computational protocol was introduced to calculate the structure of the complex directly from completely random RNA and protein starting coordinates without any ad hoc assumptions or docking steps. Although some of the technical solutions presented here may be specific to the U1A system, others may be of general validity to other protein–nucleic acid complexes as well.

Coordinates for this refined structure of the U1A protein/PIE-RNA complex have been deposited at the Brookhaven Protein Databank, under accession codes 1aud (structure) and r1audmr (restraints).

Acknowledgements

We thank Chris Oubridge for the suggestion that the Tyr^{31}His Gln^{36}Arg double mutant could have improved stability under NMR conditions, Fareed Aboul-Ela for help with aspects of the structure calculations, Kiyoshi Nagai, Jo Avis and Charles Gubser

for helpful discussions and their continuing interest in the project, and Andres Ramos and Andreas Haaf for helpful discussions. P.W.A.H. and F.H.-T.A. were supported by fellowships from the MRC, F.H.-T.A. also by Ecole Normale Superieure (Paris).

References

Allain, F.H.-T. and Varani, G. (1995) *J. Mol. Biol.*, **250**, 333–353.

Allain, F.H.-T., Gubser, C.C., Howe, P.W.A., Nagai, K., Neuhaus, D. and Varani, G. (1996) *Nature*, **380**, 646–650.

Allain, F.H.-T. and Varani, G. (1997) *J. Mol. Biol.*, **267**, 338–351.

Allain, F.H.-T., Howe, P.W.A., Neuhaus, D. and Varani, G. (1997) *EMBO J.*, **16**, 5764–5774.

Avis, J., Allain, F.H.-T., Howe, P.W.A., Varani, G., Neuhaus, D. and Nagai, K. (1996) *J. Mol. Biol.*, **257**, 398–411.

Battiste, J.L., Mao, H., Rao, N.S., Tan, R., Muhandiram, D.R., Kay, L.E., Frankel, A.D. and Williamson, J.R. (1996) *Science*, **273**, 1547–1551.

Bax, A., Ikura, M., Kay, L.E. and Zuo, G. (1991) *J. Magn. Reson.*, **91**, 174–178.

Belrhali, H., Yaremchuk, A., Tukalo, M., Larsen, K., Berthet-Colominas, C., Leberman, R., Beijer, B., Sproat, B., Als-Nielsen, J., Grübel, G., Legrand, J.-F., Lehmann, M. and Cusack, S. (1994) *Science*, **263**, 1432–1436.

Brünger, A.T. (1990) *X-PLOR Manual*, Yale University Press, New Haven, CT.

Caverelli, J., Rees, B., Ruff, M., Thierry, J.-C. and Moras, D. (1993) *Nature*, **362**, 181–184.

Clore, G.M. and Gronenborn, A.M. (1994) *Methods Enzymol.*, **239**, 349–363.

Cusack, S. (1995) *Nat. Struct. Biol.*, **2**, 824–831.

Davis, A.L., Keeler, J., Laue, E.D. and Moskau, D. (1992) *J. Magn. Reson.*, **98**, 207–216.

Diamond, R. (1995) *Acta Crystallogr.*, **D 51**, 127–135.

Fletcher, C.M., Jones, D.N.M., Diamond, R. and Neuhaus, D. (1996) *J. Biomol. NMR*, **8**, 292–310.

Gerchman, S.E., Graziano, V. and Ramakrishnan, V. (1994) *Protein Exp. Purif.*, **5**, 242–251.

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

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Gubser, C.C. and Varani, G. (1996) *Biochemistry*, **35**, 2253–2267.

Gunderson, S.I., Vagner, S., Polycarpou-Schwarz, M. and Mattaj, I.W. (1997) *Genes Dev.*, **11**, 761–773.

Hall, K.B. (1994) *Biochemistry*, **33**, 10076–10088.

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.

Hommel, U., Harvey, T.S., Driscoll, P.C. and Campbell, I.D. (1992) *J. Mol. Biol.*, **227**, 271–282.

Howe, P.W.A., Nagai, K., Neuhaus, D. and Varani, G. (1994) *EMBO J.*, **13**, 3873–3881.

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

Kanaar, R., Lee, A.L., Rudner, D.Z., Wemmer, D.E. and Rio, D.C. (1995) *EMBO J.*, **14**, 4530–4539.

Marion, D. and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun.*, **113**, 967–974.

Marion, D., Ikura, M. and Bax, A. (1989) *J. Magn. Reson.*, **84**, 425–430.

McDonald, I.K. and Thornton, J.M. (1994) *J. Mol. Biol.*, **238**, 777–793.

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

Nagai, K., Oubridge, C., Ito, N., Avis, J. and Evans, P. (1995) *Trends Biochem. Sci.*, **20**, 235–240.

Nagai, K. (1996) *Curr. Opin. Struct. Biol.*, **6**, 53–61.

Nilges, M., Gronenborn, A.M., Brünger, A.T. and Clore, G.M. (1988) *Protein Eng.*, **2**, 27–38.

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

Oubridge, C., Ito, N., Evans, P.R., Teo, C.-H. and Nagai, K. (1994) *Nature*, **372**, 432–438.

Oubridge, C., Ito, N., Teo, C.-H., Fearnley, I. and Nagai, K. (1995) *J. Mol. Biol.*, **249**, 409–423.

Price, S.R., Ito, N., Oubridge, C., Avis, J.M. and Nagai, K. (1995) *J. Mol. Biol.*, **249**, 398–408.

Price, S.R., Oubridge, C., Varani, G. and Nagai, K. (1998) In *RNA–Protein Interaction: Practical Approach* (Ed., Smith, C.), Oxford University Press, Oxford, in press.

Puglisi, J.D., Chen, L., Blanchard, S. and Frankel, A.D. (1995) *Science*, **270**, 1200–1203.

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

Rould, M.A., Perona, J.J. and Steitz, T.A. (1991) *Nature*, **352**, 213–218.

Shaka, A.J., Barker, P. and Freeman, R. (1985) *J. Magn. Reson.*, **64**, 547–552.

Shaka, A.J., Lee, C.J. and Pines, A. (1988) *J. Magn. Reson.*, **77**, 274–293.

Valegárd, K., Murray, J.B., Stockley, P.G., Stonehouse, N.J. and Liljas, L. (1994) *Nature*, **371**, 623–626.

Varani, G., Aboul-ela, F. and Allain, F.H.-T. (1996) *Prog. NMR Spectrosc.*, **29**, 51–127.

Waltho, J.P. and Cavanagh, J. (1993) *J. Magn. Reson.*, **A103**, 338–348.

Wimberly, B.T. (1992) Ph.D. Thesis, University of California, Berkeley, CA.

Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson.*, **B101**, 201–205.

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

Ye, X., Kumar, R.A. and Patel, D.J. (1995) *Chem. Biol.*, **2**, 827–840.

Ye, X., Gorin, A., Ellington, A.D. and Patel, D.J. (1996) *Nat. Struct. Biol.*, **3**, 1026–1033.