**Roberto N. De Guzman**
**Ryan B. Turner**
**Michael F. Summers**
*Howard Hughes Medical Institute, Department of Chemistry and Biochemistry, University of Maryland Baltimore County,1000 Hilltop Circle, Baltimore, MD 21250*

*Received and accepted 13 November 1998:*

# Protein–RNA Recognition

***Abstract:*** *The x-ray structure of the glutamine aminoacyl tRNA synthetase bound to its cognate tRNA$^{Gln}$ and ATP was reported by Steitz and co-workers in 1989, providing the first high resolution structure of a protein–RNA complex. Since then, high resolution structures have been reported for RNA complexes with five other tRNA synthetases, the elongation factor Tu, the bacteriophage MS2 coat protein, the human spliceosomal U1A and U2B″–U1A′ proteins, and the HIV-1 nucleocapsid protein. Although the number of high resolution structures of protein–RNA complexes are rather small, some general themes have begun to emerge regarding the nature and mechanisms of protein–RNA recognition.* © 1999 John Wiley & Sons, Inc. Biopoly 48: 181–195, 1998

***Keywords:*** *protein–RNA complexes; protein–RNA recognition; U1A protein; MS2 coat protein; U2B″ protein; synthetase–tRNA complex; HIV-1 nucleocapsid protein*

## INTRODUCTION

The recognition of specific nucleic acid sequences by proteins is essential to all known life forms. Although much has been learned about the mechanisms by which proteins recognize and bind specifically to duplex, B-form DNA, very little is known about the molecular determinants of RNA recognition by proteins. Since the first structure of a protein–RNA complex appeared almost a decade ago, only a handful of structures of protein–RNA complexes have been solved by x-ray crystallography or NMR spectroscopy (Table I).

Despite the relative paucity of structural data, some generalizations regarding the molecular mechanisms of RNA recognition can be made. In most cases examined to date, proteins bind to the single-stranded regions of RNA hairpins and loops and often induce significant conformational changes. Bases in the loop regions of RNA become splayed or extruded into protein hydrophobic cavities, including clefts on the surfaces of $\beta$-sheets, and sequence specificity is conferred by a variety of hydrophobic, electrostatic, base-stacking, and hydrogen-bonding contacts with protein residues. Binding can also have a significant effect on the protein conformation. For example, the HIV-1 nucleocapsid protein undergoes dramatic structural changes upon binding to an RNA stem–loop recognition element. In contrast, RNA binding to the MS2 coat protein occurs with large changes in the RNA conformation but without significantly affecting the

*Correspondence to:* Michael F. Summers; email: summers@hhmi.umbc.edu

Contract grant sponsor: National Institutes of Health
Contract grant number: GM42561

Biopolymers (Nucleic Acid Sciences), Vol. 48, 181–195 (1998)

CCC 0006-3525/99/020181-15

**Table I Structures of Protein–RNA Complexes**

| Protein | Source | RNA | Method | Refs. |
|---|---|---|---|---|
| GlnRS | *E. coli* | $\text{tRNA}^{\text{Gln}}$ | X-ray 2.5Å | 7, 8 |
| AspRS | Yeast | $\text{tRNA}^{\text{Asp}}$ | X-ray 3.0Å | 9,10 |
| SerRS | *T. thermophilus* | $\text{tRNA}^{\text{Ser}}$ | X-ray 2.9Å | 11,12 |
| LysRS | *T. thermophilus* | $\text{tRNA}^{\text{Lys}}$ (*E. coli*) | X-ray 2.8Å | 15 |
| ProRS | *T. thermophilus* | $\text{tRNA}^{\text{Pro}}$ | X-ray 3.5Å | 14 |
| PheRS | *T. thermophilus* | $\text{tRNA}^{\text{Phe}}$ | X-ray 3.3Å | 13 |
| EF-Tu | *T. aquaticus* | $\text{Phe-tRNA}^{\text{Phe}}$ | X-ray 2.7Å | 24 |
| MS2 coat protein | Bacteriophage MS2 | Hairpin | X-ray 2.7Å | 28,29 |
| MS2 coat protein | Bacteriophage MS2 | Aptamers | X-ray 2.8Å | 32,33 |
| U1A | Human | Hairpin II | X-ray 1.9Å | 34 |
| U1A | Human | 3′-UTR | NMR | 36,37 |
| U2B″–U2A′ | Human | Hairpin IV | X-ray 2.4Å | 41 |
| NC protein | HIV-1 | SL3 | NMR | 42 |
| Reverse transcriptase | HIV-1 | Pseudoknot | X-ray 4.8Å | 51 |

backbone conformation of the protein. In most cases, RNA recognition appears to occur via an "adaptive binding" mechanism wherein the RNA, and sometimes the protein, undergoes significant conformational changes upon complex formation.

This article reviews the high resolution structural data that have been reported for RNA complexes with intact proteins or folded protein domains. Some aspects of protein–RNA recognition have been reviewed recently.[1–4]

## AMINOACYL TRNA SYNTHETASE COMPLEXES WITH THEIR COGNATE TRNAS

The aminoacyl–tRNA synthetases (aaRS) are a diverse group of enzymes that ensure the fidelity of protein translation by attaching the correct amino acids to their cognate tRNAs. The 20 synthetases are divided into two classes (I and II) with 10 members each, based on sequence alignment and structural homology. Class I synthetases contain a five-handed parallel $\beta$-sheet, or Rossmann fold, and aminoacylate tRNAs at the terminal 2′-hydroxyl group. Class II synthetases contain three conserved motifs and aminoacylates the terminal 3′-hydroxyl group. The structures of the synthetases have been reviewed recently.[5]

In order for the synthetases to function correctly, they must be able to discriminate among the 20 amino acids and identify their cognate tRNAs.[6] The first structure of a protein–RNA complex came from this field, and to date, the structures of six synthetase-tRNA complexes have been reported—that of GlnRS,[7,8] AspRS,[9,10] SerRS,[11, 12] PheRS,[13] ProRS,[14] and LysRS.[15] Work continues in this area to eventually obtain the structures of the 20 synthetases.[2]

### Class I tRNA Synthetase–tRNA Complexes

The co-crystal structure of the *Escherichia coli* GlnRS complexed with $\text{tRNA}^{\text{Gln}}$ and ATP[7,8] (Figure 1) is the first structure of a protein–RNA complex to be determined, and remains the only example of a class I synthetase–tRNA complex structure. In the complex, $\text{tRNA}^{\text{Gln}}$ retains the overall L-shape that has become the characteristic hallmark of tRNA molecules described in biochemistry textbooks (Figure 1A). The protein binds to the inside of the L, with protein–RNA contacts occurring at the anticodon stem–loop, the D arm, and the single-stranded regions of the acceptor arm. Although the structure of free $\text{tRNA}^{\text{Gln}}$ is not available, the similarity of the bound $\text{tRNA}^{\text{Gln}}$ crystal structures of isolated $\text{tRNA}^{\text{Phe}}$,[16,17] $\text{tRNA}^{\text{Asp}}$,[18] and $\text{tRNA}^{\text{fMet}}$ [19] imply that the overall structure and folding of tRNA molecules are similar. Comparison of bound $\text{tRNA}^{\text{Gln}}$ with that of free tRNAs indicated that the most significant structural changes occur in the anticodon loop and the acceptor arm upon binding to the protein. In the free tRNAs, the anticodon bases are base stacked within the loop. In the complex, the anticodon bases (C34, U35, G36) are extruded from the loop and interact with hydrophobic pockets on the protein surface. The complex is further stabilized by specific intermolecular hydrogen bonds (Figure 1B). All the possible hydrogen-bond donors and acceptors in the loop bases, except that of G36 N3, are involved in forming hydrogen bonds with the protein.

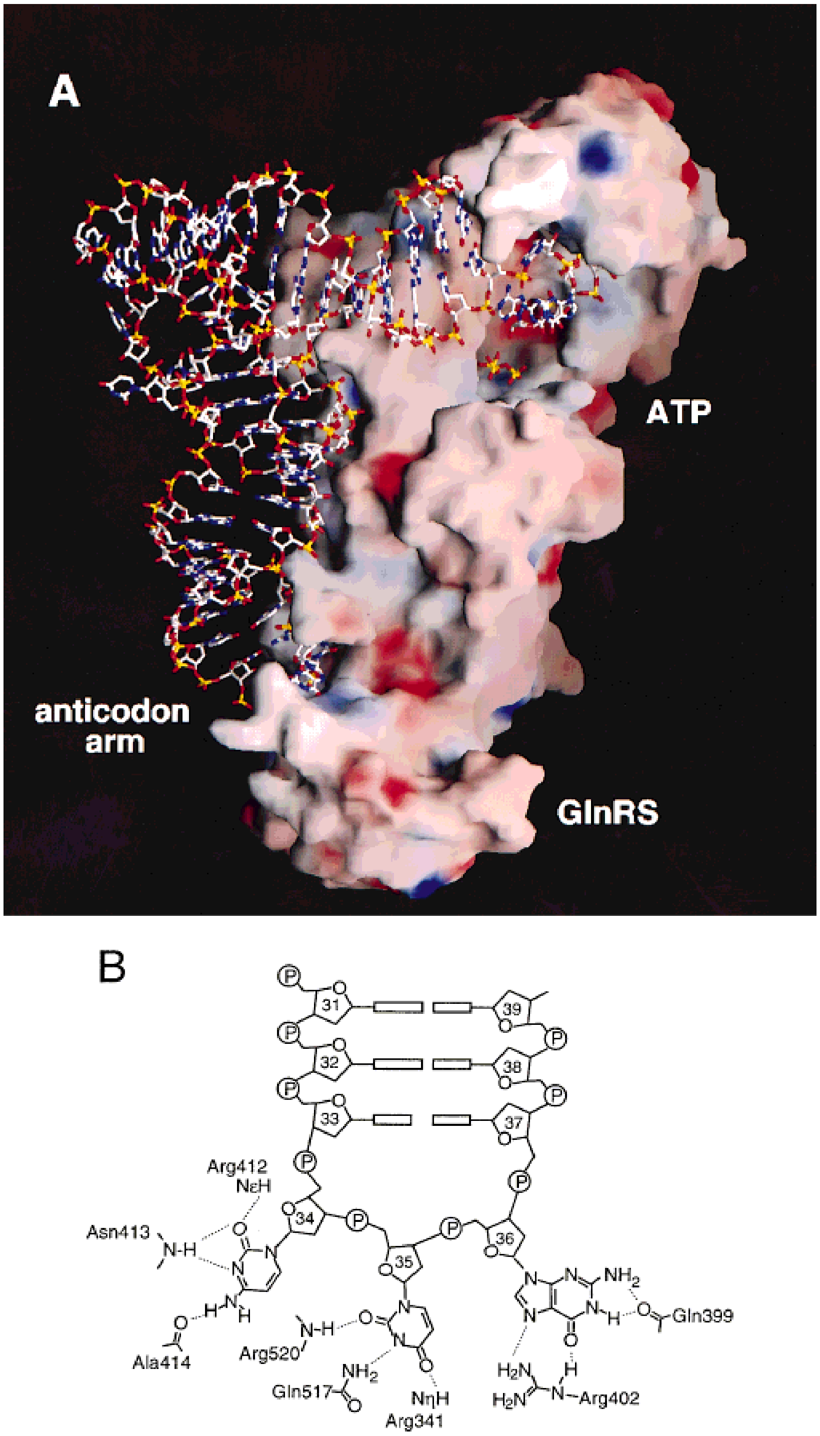


**FIGURE 1** (A) Structure of *E. coli* GlnRS complexed with tRNA$^{Gln}$ and ATP.[7, 8] The protein is depicted as a surface representation and the tRNA as a stick model. The acceptor arm of tRNA$^{Gln}$ sits deep in a protein pocket that is very close to the ATP. All surface representations are generated from the available PDB coordinates by GRASP.[52] (B) The anticodon bases are splayed into the protein and form an extensive network of hydrogen bonding interaction with protein residues.

The identity elements of *E. coli* tRNA$^{Gln}$ are the anticodon bases (C34, U35, G36), the first three base pairs (U1–A72, G2–C71, G3–C70) and G73 in the acceptor stem.[20,21] The co-crystal structure confirms the roles of these bases in tRNA recognition. GlnRS binds to the minor groove of the acceptor stem and disrupts the first base pair (U1-A72). U1 is disordered in the co-crystal and A72 participates in a hydrophobic interaction with the side chain of Leu136. The base pairs G2–C71 and G3–C70 are recognized by base-specific hydrogen bonds formed by the amino group of G2 and G3 and the exocyclic oxygen atom of C71 with protein residues. The single-stranded region of the acceptor stem is buried deep in a protein cavity and lies close to ATP and the binding site for glutamine. G73, C75 , and C76 are base stacked to each other, with the amino group of G73 forming an intramolecular hydrogen bond with the phosphate of A72. The base of C74 projects in a protein hydrophobic pocket and the 2-amino group forms hydrogen bonds with backbone carbonyl groups. The four modified bases in the anticodon arm form non-Watson-Crick base pairs, two of which (residues Φ38 and 2-mehtyl-A37) also form hydrogen bonds with the side chain of Asn370.

## Class II tRNA Synthetase–tRNA Complexes

The co-crystal structure of the yeast AspRS bound to tRNA$^{Asp}$,[9,10] determined at 2.9 Å resolution, represents an example of a class II synthetase–tRNA complex. The homodimeric AspRS binds two tRNA$^{Asp}$ molecules at the anticodon and acceptor arms (Figure 2). Protein binding induced a large conformational change in the anticodon loop of the tRNA molecule. Five bases in the anticodon loop, from residues 33 to 37, are splayed and interact in a protein surface formed by a five-stranded $\beta$-barrel (Figure 2B). Protein residues from the surface of the $\beta$-strands contact the anticodon bases and form hydrophobic and hydrogen-bonding interactions, and the three anticodon bases (G34, U35, and C36) form base-specific hydrogen bonds (Figure 2C). The modified base pair, Ψ32–C38, that closes the anticodon loop is also recognized by the protein via intermolecular hydrogen bonds to the Ψ32 N3 and C38 N4 and N3 atoms. In contrast to GlnRS, AspRS binds to the major groove of the acceptor stem and does not disrupt the first base pair (U1–A72). This approach from the major groove results in an opposite orientation of the CCA end of tRNA$^{Asp}$ as compared to tRNA$^{Gln}$ (for comparison, see Figures 1A and 2A). The protein also forms base-specific hydrogen bonds with G73 and the first base pair U1–A72. The identity elements of yeast tRNA$^{Asp}$ include the anticodon bases G34, U35, C36, the discriminator base G73 in the acceptor arm, the G10–U25 base pair,[22] and C38[23] in the D arm. The co-crystal structure showed direct base-specific protein contacts to all these determinants, except the G10–U25 base pair. The protein also interacts with the ribose and phosphate groups of neighboring residues U11 and U12.

GlnRS and AspRS are examples of synthetases that recognize the anticodon loops of tRNAs. The splaying of the anticodon bases is also seen in co-crystal structures from the class II synthetases from *Thermus thermophilus* LysRS[15] and ProRS.[14] In these structures, the synthetases approach the anticodon loops from the major-groove side. However, in the co-crystal structure of PheRS,[13] the protein approaches the anticodon loop from the minor-groove side. In addition, the anticodon bases in the PheRS–RNA$^{Phe}$ complex are not splayed into the protein, but instead retain the conformation observed in the structure of the free tRNA$^{Phe}$. Thus, even for a given class of RNA-binding proteins, the mechanism of recognition and binding can be dramatically different.

The crystal structure of the *T. thermophilus* SerRS–tRNA$^{Ser}$ complex was determined at 2.7 Å resolution[11,12] (Figure 3). The tRNA binds across the two subunits of the homodimeric SerRS, resulting in a reorientation of a coiled-coil domain that packs between the TΨC and the variable arms of tRNA$^{Ser}$. Aside from this reorientation, the overall structures of the free and bound SerRS remain relatively similar. A major difference between the SerRS and the synthetases discussed above is the mode of recognition. This complex offers an example of synthetase–tRNA recognition that is not based on the anticodon arm, but instead involves the phosphate and ribose backbone of the tRNA. In fact, the anticodon arm, from residues C27 to G41, is disordered in the co-crystal. Unlike the other synthetases, the co-crystal structure of SerRS shows only three base-specific hydrogen-bonding contacts involving G19 in the D loop and the G47a–C47n base pair in the variable stem. The amino group of G19 forms a hydrogen bond with backbone carbonyl group of Ala555, and the side chain of Gln545 forms hydrogen-bonding contacts with G47a N2H and C47n O2. One distinguishing feature of tRNA$^{Ser}$ is the presence of a long variable arm. In the co-crystal structure, most of the protein–RNA contacts occur along the backbone phosphates located in the long variable arm. The variable loop (C47e to A47j) is also disordered in the co-crystal and do not participate in protein binding. The protein also contacts the phosphates of the acceptor stem and the TΨC loops. No structural information is available for the single-stranded region of the acceptor arm because the ends of the acceptor stem (G1–A3 and C72–A76) are also disordered in the co-crystal.

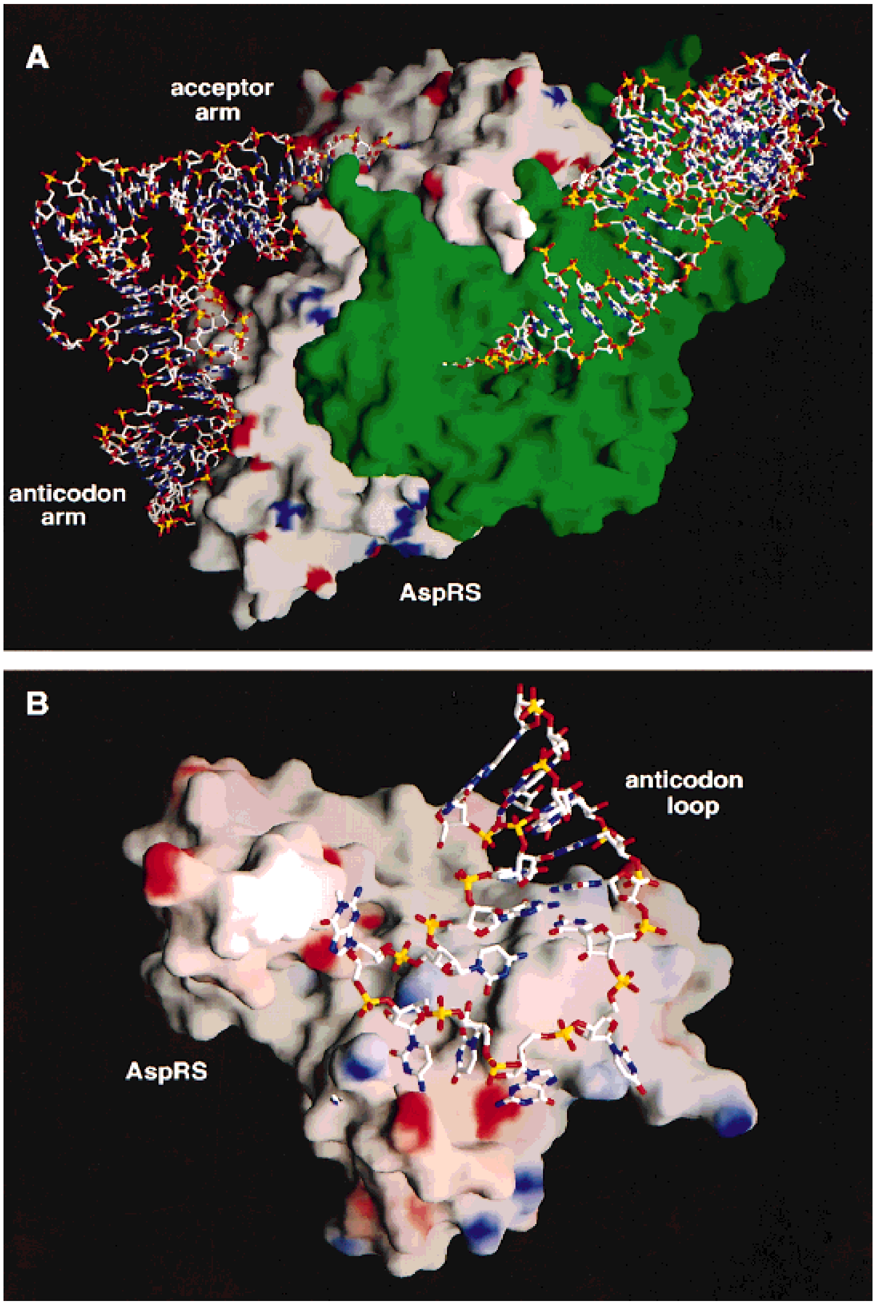


**FIGURE 2** (A) Structure of yeast AspRS bound to tRNA$^{Asp}$ and ATP.[9,10] One protein monomer is colored green. The AspRS dimer binds two tRNA molecules. (B) Binding of the anticodon loop of tRNAAsp to AspRS. The anticodon bases are splayed and interact with the protein. (C) Details of the extensive network of protein–RNA hydrogen bonds formed by tRNA$^{Asp}$ with AspRS.

## EF-TU–TRNA COMPLEX

During translation in prokaryotes, the aminoacylated tRNA (aa-tRNA) is recruited into ribosomes by the elongation factor Tu (EF-Tu). EF-Tu is a GTP-binding protein and binds to aa-tRNA when complexed with GTP and releases the tRNA when GTP is hydrolyzed to GDP. Unlike the synthetases, where each synthetase must recognize only its cognate tRNA, EF-Tu must be able to recognize all the aa-tRNAs. The structure of the ternary complex formed by *Thermus aquaticus* EF-Tu, yeast Phe-tRNA$^{Phe}$, and the GTP analogue GDPNP[24] reveals the structural basis for the recognition of aminoacylated tRNA by EF-Tu

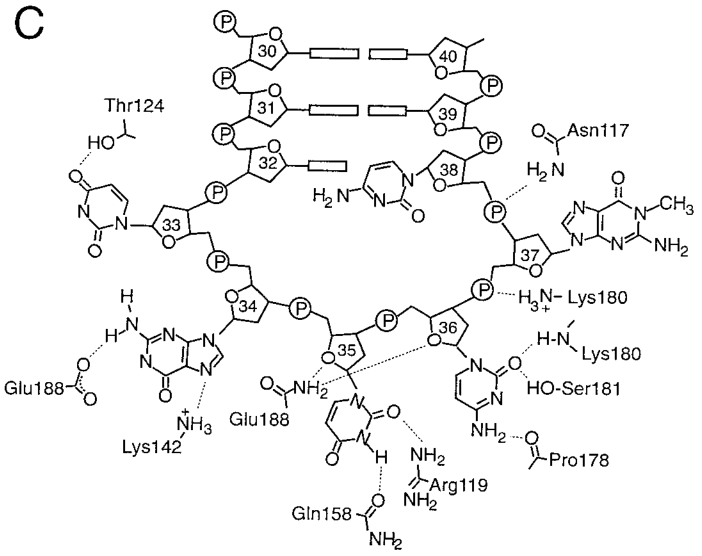


**FIGURE 2** (*Continued*)

(Figure 4). The protein binds only one end of the aa-tRNA, leaving the rest of the tRNA free to interact with the ribosomes. In this manner, the anticodon loop can still form base pairs with the mRNA in the ribosome during the elongation phase of protein synthesis. The phenylalanine at the end of the tRNA$^{Phe}$ complex

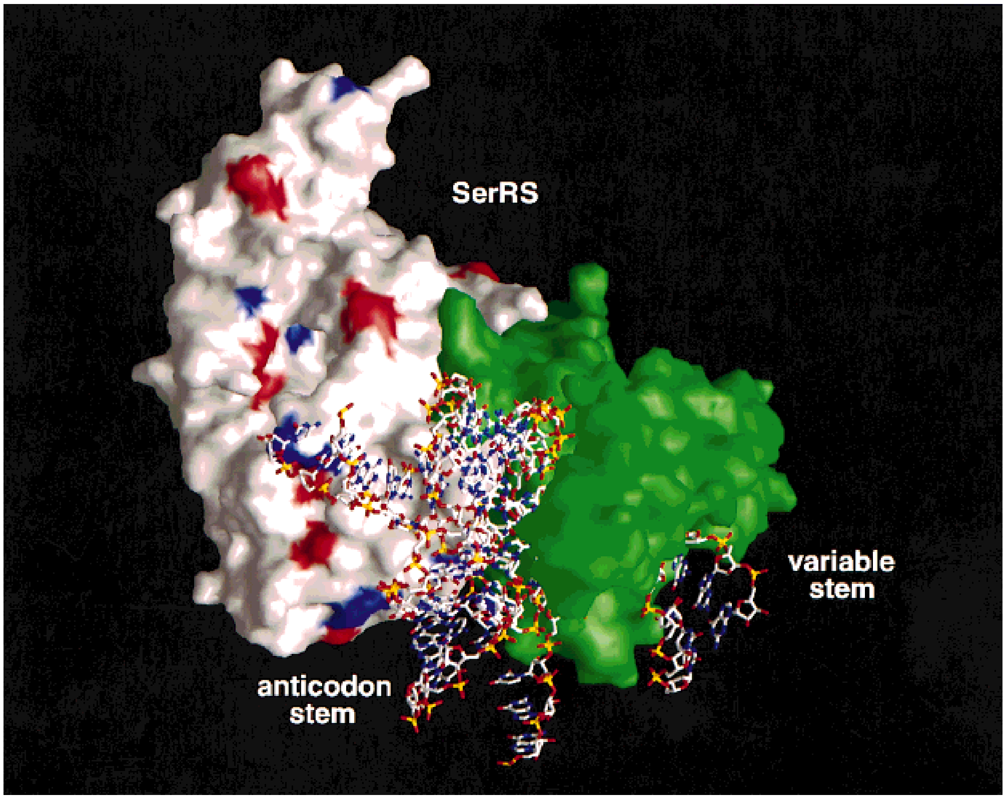


**FIGURE 3** Structure of *Thermus thermophilus* SerRS bound to tRNA$^{Ser}$ .[11,12] The tRNA$^{Ser}$ contacts both subunits of the dimeric SerRS (one monomer is colored green). A large part of the anticodon and the variable stem-loops are disordered in the co-crystal.

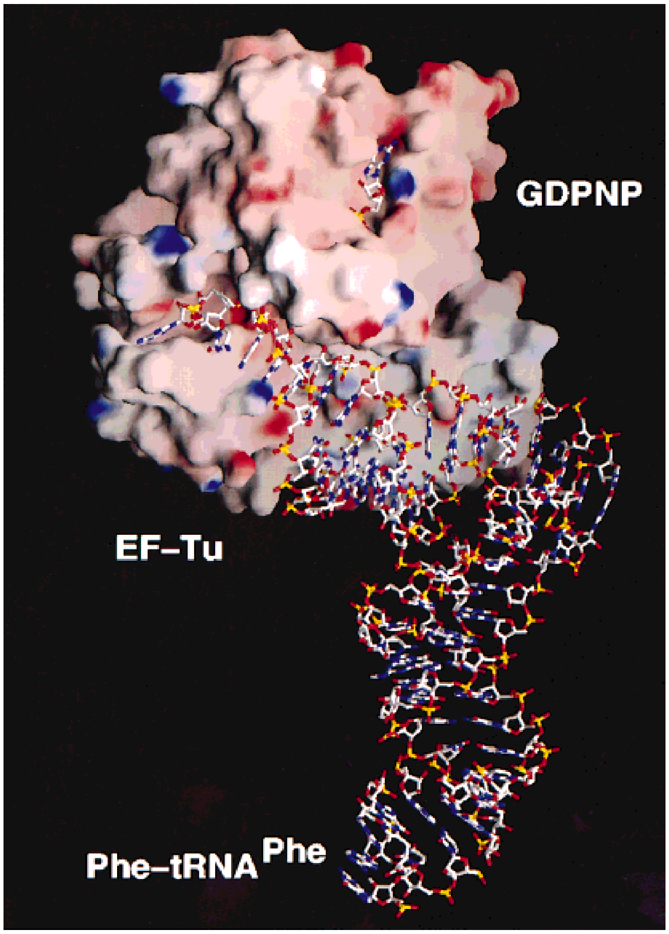


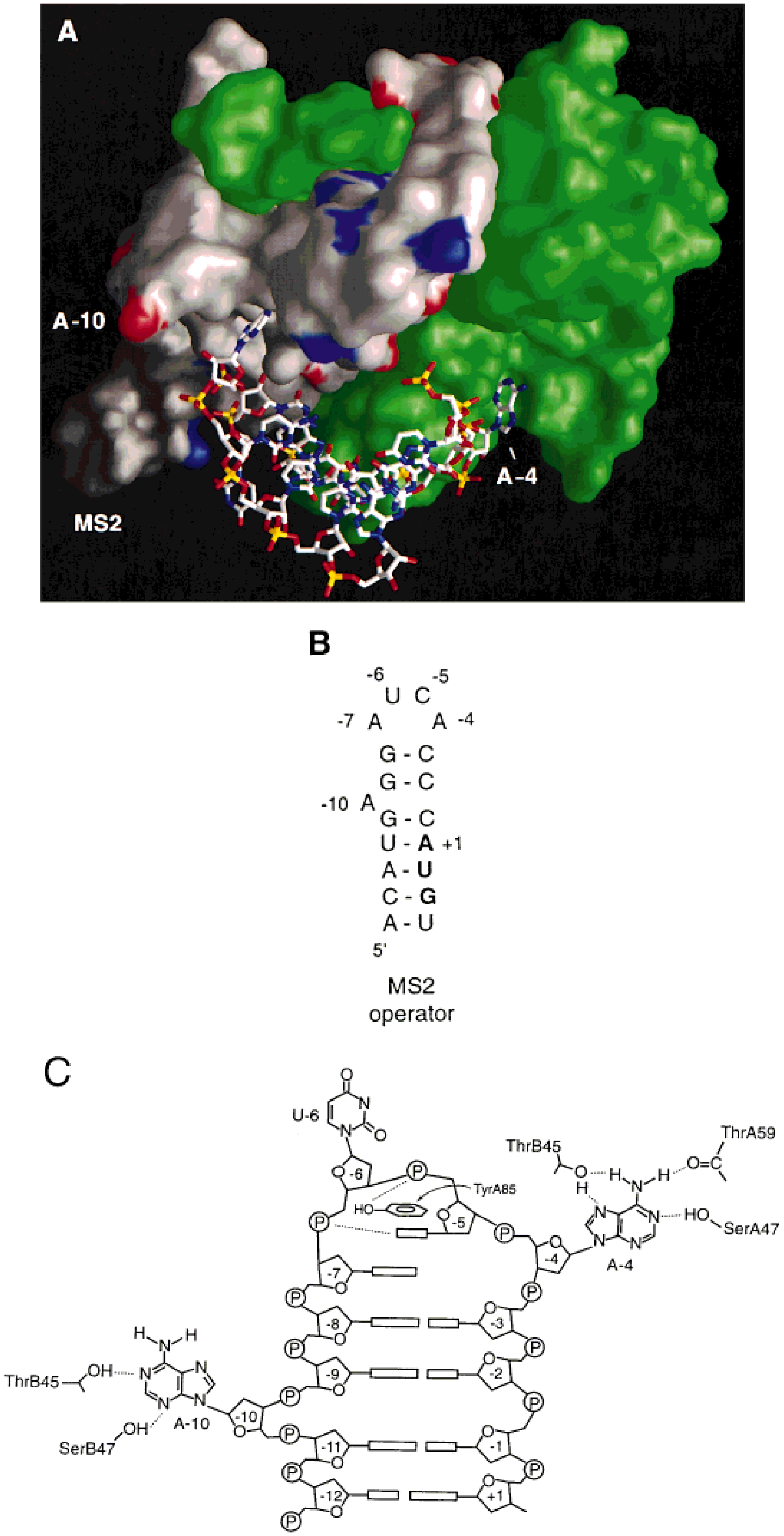
**FIGURE 4** Co-crystal structure formed by the *Thermus aquaticus* EF-Tu, yeast Phe-tRNA$^{Phe}$, and the GTP analogue GDPNP.[24] The protein binds to the acceptor arm of the tRNA.

fits in a pocket large enough to accommodate the other amino acids (Figure 4) and the amino ester link forms hydrogen bonds with protein backbone carbonyl and amide groups. The three terminal bases, CCA–Phe, are internally stacked and make protein contacts via the phosphate backbone. The protein also contacts the phosphate backbone of G1 and C2 and the ribose of C2. The protein also makes contacts with the ribose and phosphate groups of the TΨC stem of the tRNA.

## MS2 COAT PROTEIN–HAIRPIN RNA COMPLEX

MS2 is a small single-stranded RNA bacteriophage of *E. coli*. MS2 coat proteins form dimers at low concentrations (1 n*M*–1$\mu$*M*) and aggregate into stable phage-like capsids at higher concentrations.[25] The crystal structure of the empty capsid, refined to 2.8 Å resolution,[26,27] reveals that the capsid comprises 180 coat proteins arranged in an icosahedral T=3 symmetry. Two types of dimers (A/B and C/C) in the capsid are held together by an extensive network of hydrophobic contacts at the dimer interface.

The dimer binds a unique hairpin in the MS2 genome. RNA binding is required for specific encapsidation of the MS2 genome, and since the hairpin contains the initiation codon of the replicase gene, binding represses the translation of the viral replicase gene. The co-crystal was obtained by soaking the 19-nucleotide RNA hairpin into the MS2 capsids and the structure of the MS2 coat protein–hairpin RNA was determined at 2.7 Å resolution, Figure 5.[28, 29] The RNA binds to both subunits of the dimer and does not alter the overall structure of the protein. However, complex formation results in a large conformational change in the RNA hairpin. Nuclear magnetic reso-


A
A-10
A-4
MS2
B
-6 U
C -5
-7 A
A -4
G - C
G - C
-10 A
G - C
U - A +1
A - U
C - G
A - U
5'
MS2
operator
C
U-6
ThrB45
ThrA59
TyrA85
HO
SerA47
A-4
-6
-5
-4
-7
-8
-9
-10
-11
-12
-3
-2
-1
+1
P
ThrB45
OH
SerB47
OH
A-10
H
N
H

nance data suggest that the bulged adenosine at position -10 is stacked within the stem in the free RNA.[30] Upon binding to the protein, the base of A-10 is flipped outside the stem and binds to a hydrophobic pocket formed by the side chains of ValB29 and LysB61 (each protein subunit labeled with A and B), with hydrogen bonds from N1 and N3 to the hydroxyl groups of ThrB45 and SerB47, respectively (Figure 5C). A point mutation (Thr45 → Ala) introduced to disrupt this hydrogen bond decreased the affinity but did not completely eliminate binding.[31] In the loop region, the base of A-4 is splayed into a protein hydrophobic pocket lined with the side chains of ValA29 and LysA61 and the 4-amino, N1, and N7 groups of A-4 form hydrogen bonds with protein residues.

Biochemical studies indicate that the A-7, A-10, and A-4 positions are needed for binding, and that the -5 position can be either a cytosine or a uracil.[25] Substitution of cytosine for the wild-type uracil at position -5 increased the binding affinity by 100-fold. The co-crystal structure showed that the exocyclic amino group of C-5 forms a hydrogen bond with the phosphate at -6 position. Recent co-crystal structures of the MS2 coat protein with three different RNA aptamers reveal that, despite differences in the RNA sequences and structures, the aptamers bind to the same sites on the inner surface of the capsid.[32,33]

## SPLICEOSOMAL PROTEIN–RNA COMPLEXES

Pre-mRNA splicing occurs in a large multicomponent assembly called the spliceosome, which is built up from several small nuclear ribonucleoproteins (snRNPs). Each snRNP, in turn, is made up of several proteins in association with a small nuclear RNA (snRNA) molecule. One of the protein components of the human U1 snRNP is the U1A protein, a 282-residue protein that binds to the 164-nucleotide U1 snRNA. Only the N-terminal third of the U1A protein, from residues 2 to ~ 100, is needed for RNA binding, and this part is referred to as the ribonucleoprotein (RNP) domain, the RNA-binding domain (RBD), or the RNA recognition motif (RRM). The RNP domain of U1A recognizes the sequence, AUUGCAC, located in hairpin II of the U1 snRNA and the polyadenylation signal in the 3′-untranslated region (3′-UTR) of the U1 mRNA. The RNA binding property of U1A plays a dual role—(a) binding to hairpin II stabilizes the folding of the U1 snRNA, and (b) binding to the 3′-UTR prevents polyadenylation and regulates the translation of the U1A protein.

## U1A–HAIRPIN II COMPLEX

The crystal structure of the U1A RNP domain (residues 1–98) complexed with the 21-nucleotide hairpin II RNA[34] was determined at 1.9 Å resolution (Figure 6A).Protein–RNA recognition occurs in the single-stranded region of hairpin II. The protein forms a hydrophobic channel that accommodates the loop bases. The highly structured $\beta2$–$\beta3$ loop protrudes through the 10-nucleotide RNA loop, and the first seven nucleotides fit into the groove between the $\beta2$–$\beta3$ loop and the C-terminal region of the protein. The last three nucleotides, U[13]CC, are found to have no contacts with the protein and are much less ordered. The base of U7 is stacked with A6, which in turn is stacked with G5. The bases from U8 to C12 are splayed into the protein hydrophobic channel.

The C5–G16 base pair that closes the loop plays an integral role in positioning the RNA for contact with the protein. This protein–RNA contact is mediated by hydrogen bonds with Arg52 to the N7 and O6 of G16 and with the N1 of A6. Mutation of Arg52 to Gln completely abolishes RNA binding. Arg52 is aptly located at the $\beta2$–$\beta3$ loop, where recognition of the G16 and A6 facilitates insertion of U1A through the RNA loop. This insertion results in a further opening of the loop, which essentially presents a single-stranded nucleic acid for binding to the protein. Residues A6 and U7, representing the first two single-stranded nucleotides, are base stacked and continue the helical geometry of the RNA stem. U8 partially stacks with the purine ring of G9, and G9 is packed against the side chain of Gln54. The pyrimidine ring of C10 stacks on the aromatic side chain of Tyr13, and a series of stacking interactions between the phenyl ring of Phe56, the bases of A11 and C12, and the side chain of Asp92 constitute a major recognition

**FIGURE 5** (A) Surface representation of the bacteriophage MS2 coat protein bound to a hairpin RNA.[28, 29] The protein monomers are colored gray and green with blue basic residues and red acidic residues. (B) Sequence of the hairpin RNA used in the co-crystallization. The adenine bases in the stem (A-10) and the loop (A-4) bind into protein hydrophobic pockets and form (C) hydrogen bonds with protein residues (figure adapted from Ref. 3).

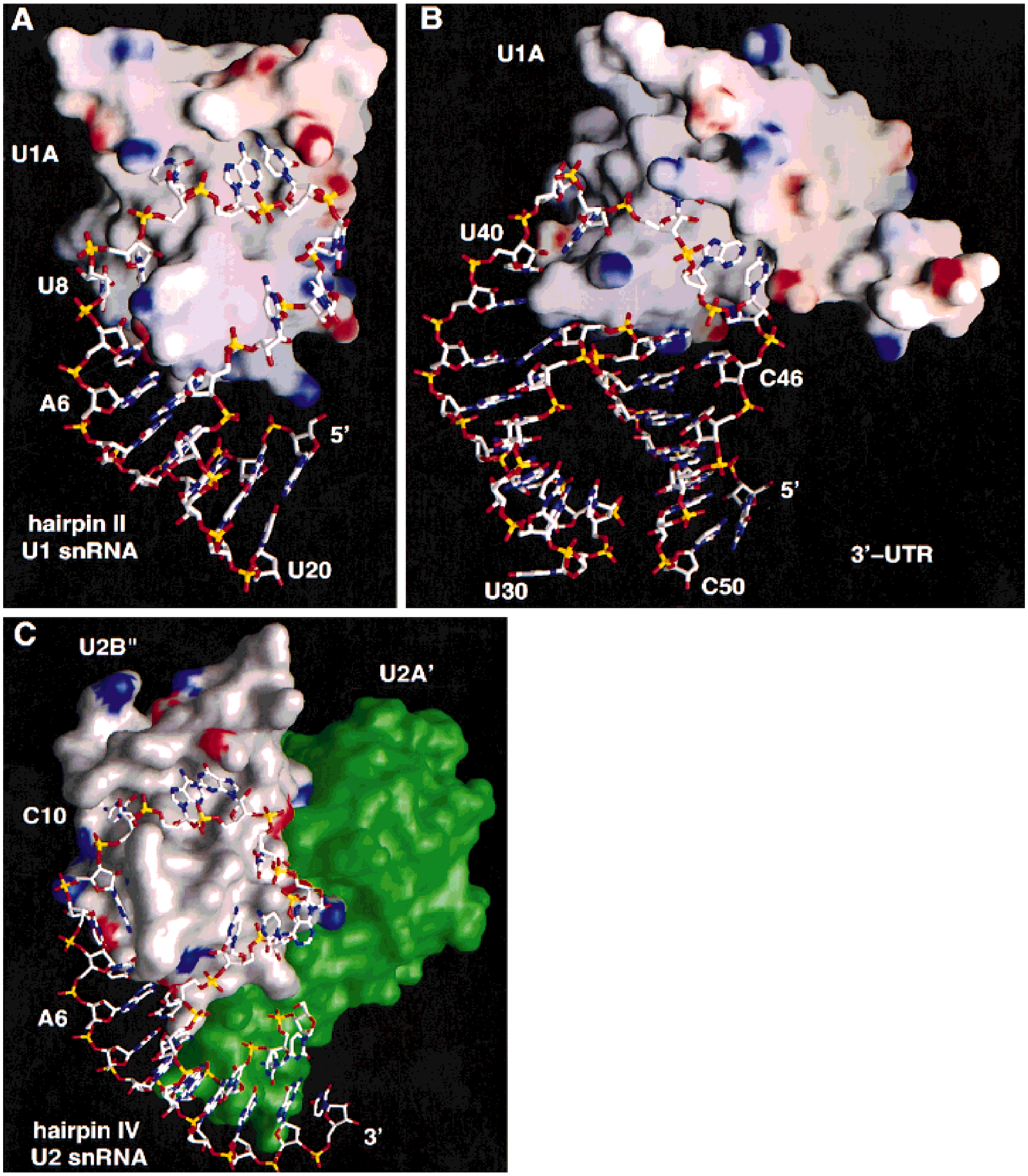


**FIGURE 6** Structures of human spliceosomal protein–RNA complexes. U1A binds to (A) hairpin II of the U1 snRNA,[34] and (B) an internal loop in the polyadenylation signal in the 3′-UTR of the U1 mRNA.[36] The tetraloop $U^{29}$UCG was used to close the stem in the 3′-UTR construct, a skip in the numbering was used to match the numbering used in the wild-type sequence. (C) Ternary complex formed by the U2B″–U2A′ and hairpin IV of U2 snRNA. U2B″binds hairpin IV only when complexed with U2A′. (D, E) RNA constructs used in structure determination of U1A protein and the (F) U2B″–U2A′complex. The U1A binds to the heptanucleotide AUUGCAC (in bold) in both constructs, and U2B″ to AUUGCAG.

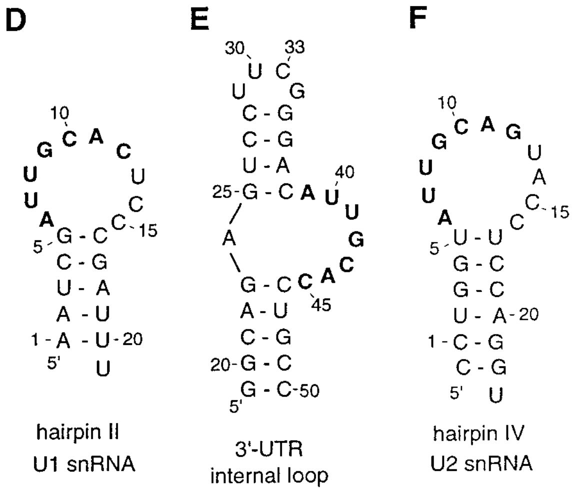


**FIGURE 6** (*Continued*)

element. A Tyr13 → Phe mutation abolishes RNA binding[35] through disruption of an extensive hydrogen-bonding network that orients the phenyl ring relative to C10.

## U1A-3′–UTR COMPLEX.

The structure of the RNP domain of human U1A protein (residues 2–102) bound to a 30-nucleotide RNA construct was solved by nmr spectroscopy[36,37] (Fig. 6B). This represents the first structure of a protein–RNA complex determined by nmr methods. Again, recognition of the RNA occurs at the AUUGCAC sequence of the internal loop of the 3′-UTR. As in the U1A–hairpin II complex, the $\beta$2–$\beta$3 loop of the U1A protein protrudes through the single-stranded region of the RNA. There is a severe kink in the RNA, and the bases of the single-stranded nucleotides are splayed into the surface of the $\beta$-sheet forming the necessary intra- and/or intermolecular stacking interactions as previously determined in the U1A–hairpin II complex.

The intermolecular interactions of residues A39–C45 are similar to those observed in the U1A–hairpin II complex, but additional interactions involving the $\beta$2–$\beta$3 and $\beta$1-helix A loops, A24, and the loop-closing residues G23 and C46 are observed. A24 interacts with Ser48 and stacks on the G23–C46 base pair, and these three nucleotides subsequently form interactions with Arg47 and Lys23. A39 stacks on U40 and on the loop-closing base pair G25–C38 in a manner similar to the above; Arg52 forms hydrogen-bonding interactions and extensive hydrophobic packing of the Leu49 side chain against the nucleotides is observed. The second interface of the RNA protein contact involves the $\beta$-sheet surface and the $\beta$4-helix C loop and the nucleotides U41–C45. This region is important because it contains three highly conserved amino acids that lead to the following stacking interactions: G24 on Gln54, C43 on Tyr13, and A44 on Phe56. The high conservation of these amino acids among the RNP family of proteins indicates the above interaction may be a general pattern of RNP–RNA recognition.

The C-terminal region of the RNP domain is crucial for recognition of RNA, including U1A,[34,35,38] and heterogeneous nuclear RNP C[39] and U1 70 K.[40] Residues 92–98 form a helix in the U1A protein, and bury highly conserved hydrophobic residues of the $\beta$-sheet in the free protein. The side chains of Ile93, Ile94, and Met97 form a hydrophobic core with the surface of the $\beta$-sheet, yet the surface of the $\beta$-sheet must be exposed for interaction with the RNA. In binding to the RNA, this helix C rotates away from the $\beta$-sheet like a “cat-flap,”[36] allowing the formation of an alternative hydrophobic core with helix C. In addition, the $\beta$-sheet and the $\beta$4-helix C loop become exposed for interaction with RNA.

## U2B″–U2A′–HAIRPIN IV COMPLEX

The U2B″–U2A′ protein complex with the 24-nucleotide hairpin IV of U2 snRNA (Figure 6C) was also determined by x-ray crystallography to 2.4 Å resolution.[41] U2B″ binds to its cognate RNA only when complexed to U2A′, which contains leucine-rich repeats (LRR). The concave surface created by the parallel $\beta$-sheet of the LRR region of U2A′ cradles helix A of U2B″ while the N- and C-terminal arms of the U2A′ LRR domain complete the interaction by grasping the U2B″ RNP domain. Hairpin IV of the U2 snRNA binds to the U2B″ in nearly the same orientation as the hairpin II–U1A complex, likely due to the high degree of conservation among the RNAs and proteins involved. The first six nucleotides of hairpin IV of U2 snRNA and hairpin II of U1 snRNA are identical, differing only in the last base, (Figure 6D and F). The U2B″ and U1A proteins differ in their residues at or near sites of RNA contact, which allows discrimination between the respective hairpins. Residues near the loop closing base pair in U1A, Leu17, and Glu19 are replaced by Met and Asp in the U2B″ protein. Six sequential residues differ at the end of the $\beta$2 strand between the U1A and U2B″ protein and these residues are in contact with the 3′ half of hairpin IV.

The C10 nucleobase stacks with Tyr13, and Phe56, A11, G12 (see Figure 6), and the side chain of Asp92 form a continuous stacking interaction in the ternary

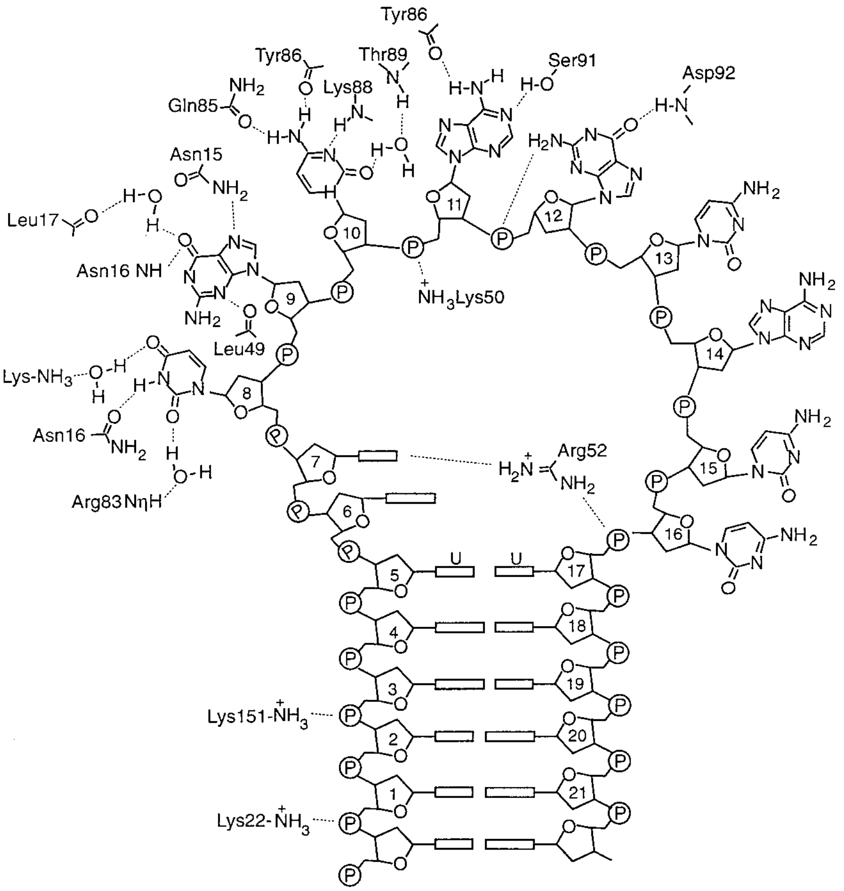


**FIGURE 7** Details of the hydrogen-bonding network formed by hairpin IV with the U2B″–U2A′ complex.

U2B″–U2A′–hairpin IV complex. Similar interactions are observed in the U1A–hairpin II complex. Interestingly, the 2-amino group of G12 is hydrogen bonded to its phosphate (Figure 7), fixing it in the *syn* conformation and permitting the larger purine base of G12 to be inserted into the same space as C12 in the hairpin II–U1A complex. The second recognition element constitutes an important difference from the hairpin II–U1A complex. In the complex, the last three nucleotides of hairpin II ($U^{13}CC$) are poorly ordered,[34] extending into solution without observable contacts with the protein, whereas the last four nucleotides of hairpin IV form a rigid structure that resembles a stepladder.[41] This localized structure is stabilized by stacking interactions between A14, C15, and C16, and by a hydrogen-bonding network involving the phosphates of C15 and C16 and the U13 base. The C16 and C13 bases pack against the side chains of Thr48 and Leu46. In the U1A protein, each of these amino acids is replaced by a serine residue, and the reduced affinity of the U1A protein for hairpin IV appears to be due to a destabilization of the stepladder structure.[41]

A third recognition element involves interactions between protein and the loop-closing base pair, U5–U17. A non-Watson–Crick U–U base pair closes the loop of hairpin IV, whereas a normal G–C pairing closes the loop of hairpin II. The much closer pairing of the backbone atoms in the U–U base pair (C1′–C1′ distance 8.8 Å as opposed to 10.5 Å in hairpin II) causes A6 to stack against U17 on the opposite strand rather than the adjacent loop-closing base, which was observed in the hairpin II–U1A complex. Arg52 of the U2B″ protein forms a salt bridge with the U17 phosphate and the carboxyl group of Asp19. For comparison, in the U1A complex, Arg52 hydrogen bonds with the O6 and N7 of G16 and the N1 of A6. Thus, in both cases, Arg52 appears to function by forming hydrogen bonds to the base pair that closes the loop (G-C in hairpin II and U-U in hairpin IV). The interaction of the RNA stem and the protein is the fourth important interaction . The LRR motif facilitates the binding of U2A′ with the RNA stem, and is important for the formation of the ternary complex. Lys151 of the U2A′ LRR and Lys22 of the U2B″ protein form salt bridges with G3 and C1, respectively.

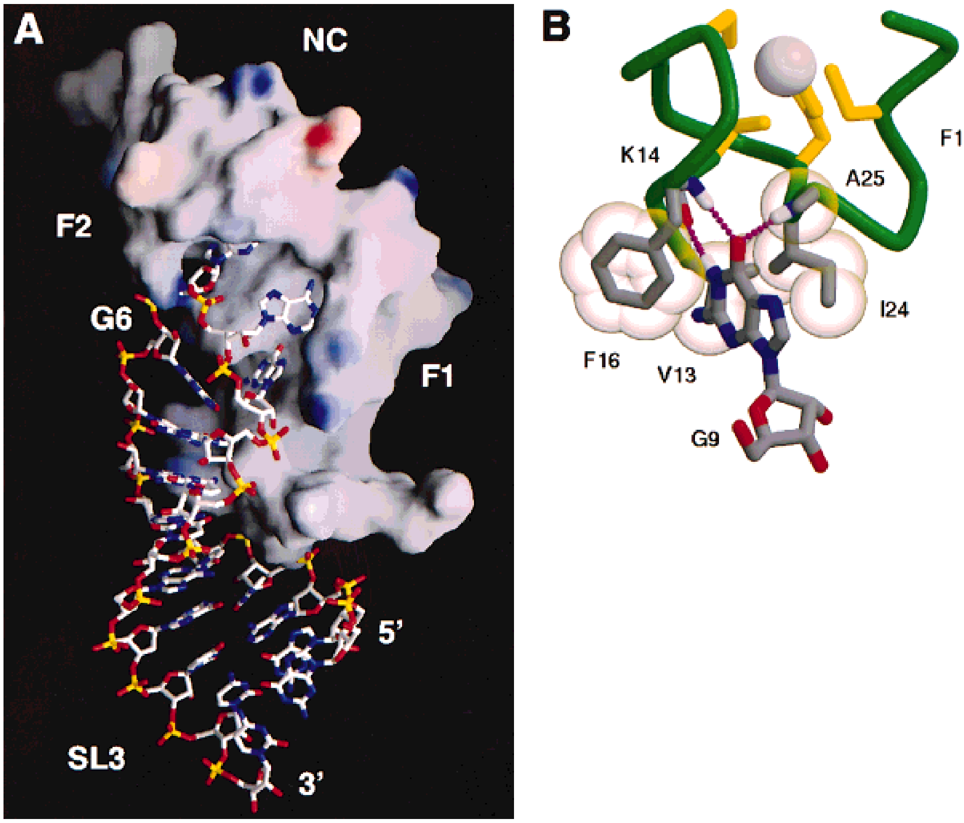


**FIGURE 8** (A) Structure of the HIV-1 NC protein bound to the SL3 stem–loop RNA. The N-terminal zinc knuckle (F1) binds to the loop base G9, and the C-terminal zinc knuckle (F2) binds G7. The N-terminal tail forms a $3_{10}$-helix that packs in the major groove of the RNA stem. (B) The N-terminal zinc knuckle F1 recognizes G9 by hydrophobic and hydrogen bonding interactions.

## HIV-1 NUCLEOCAPSID PROTEIN–SL3 RNA COMPLEX

The nmr structure of the HIV-1 nucleocapsid (NC) protein bound to the SL3 stem–loop RNA[42] represents an example of RNA recognition by retroviral zinc knuckle domains. In the mature virus, NC is a 55-residue protein with two CCHC-type zinc knuckle domains, F1 and F2, separated by a short linker region R[29]APRKK, and flanked by N- and C-terminal tails. In the free protein, only the zinc knuckle domains are structured,[43–45] and the linker and tails are flexible.[46]

SL3 is one of four stem–loops located in the packaging signal, or the Ψ site, of the HIV-1 genome. The nmr structure of free SL3 RNA showed that the RNA stem is an A-helical conformation but the loop bases do not adopt a regular stable folding.[47] In contrast, upon complex formation with SL3, the zinc knuckles bind to two of the loop guanosines (G7 and G9; Figure 8). As observed previously for an isolated zinc knuckle–oligo-DNA complex,[48] the guanosine nucleobases bind within hydrophobic clefts on the surface of the zinc knuckles. Specificity for guanosines is conferred by hydrogen bonds from the guanosine O6 and N1H atoms to backbone amide and carbonyl groups of the zinc knuckle domains (Figure 8B). The adenine base in the tetraloop forms a hydrogen bond to the side chain NϵH of Arg32. RNA binding does not change the folding of the individual knuckles but resulted in a dramatic change in the overall topology of the protein. Thus, the linker that connects the two zinc knuckles becomes ordered, the two knuckle domains pack tightly together, and the N-terminal tail forms a $3_{10}$-helix that packs within the major groove of the A-helical RNA stem.

## CONCLUSIONS

Although it has been nearly a decade since the first structure of a protein–RNA complex appeared, much remains to be learned regarding the determinants of

RNA recognition. One might have expected to find common binding modes among closely related systems, such as the aminoacyl tRNA synthetases. However, examination of the six x-ray structures currently available reveals that binding among this related group of complexes occurs via a diverse and even dramatically different subset of interactions. In some cases, such as for the tRNA$^{Gln}$ aminoacyl synthetase, the recognition portion of the RNA unstacks, affording a splayed single stranded segment that binds to an extended protein surface. However, for other cases such as the tRNA$^{Phe}$ aminoacyl synthetase, the RNA is be relatively unperturbed upon binding, with recognition mediated to a significant extent by electrostatic interactions. Proteins can also bind to double-helical regions of RNA and recognize the backbone phosphate and ribose groups of RNA by electrostatic and hydrogen bonds, as seen in the EF-Tu and SerRS complexes. In addition, the recent nmr structure of the HIV-1 nucleocapsid protein complex with the SL3 stem loop recognition element provides an example of a dramatically different mode of binding, where a largely unstructured protein condenses on the RNA and forms extensive protein–RNA and protein–protein interactions. This diversity contrasts with the structures generally involved in protein–DNA recognition, in which sequence-specific recognition occurs in major groove of duplex, B-like DNA.[4] The development of new nmr methodologies for the study of isotopically labeled RNA,[49] as well as new methods for crystallizing unusual RNA molecules and protein–RNA complexes,[50] will likely lead to the identification of a variety of new and interesting protein–RNA recognition motifs in the near future, and this should lead to a more definitive identification of the general determinants of RNA recognition by proteins.

Support from the National Institutes of Health (GM42561 to MFS) is gratefully acknowledged. RBT is supported by a UMBC Meyerhoff Undergraduate scholarship.

## REFERENCES

1. Varani, G.; Nagai, K. Annu Rev Biophys Biomol Struct 1998, 27, 407–445.
2. Uhlenbeck, O. C.; Pardi, A.; Feigon, J. Cell 1997, 90, 833–840.
3. Nagai, K. Curr Opin Struct Biol 1996,6, 53–61.
4. Steitz, T. A. In The RNA World; Gesteland, R. F., Atkins, J. F., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1993; pp 219–237,
5. Cusack, S. Nature Struct Biol 1995, 2, 824–831.
6. Saks, M. E.; Sampson, J. R.; Abelson, J. N. Science 1994, 263, 191–197.
7. Rould, M. A.; Perona, J. J.; Soll, D.; Steitz, T. A. Science 1989, 246, 1135–1142.
8. Rould, M. A.; Perona, J. J.; Steitz, T. A. Nature 1991,352, 213–218.
9. Cavarelli, J.; Rees, B.; Ruff, M.; Thierry, J. C.; Moras, D. Nature 1993, 362, 181–184.
10. Ruff, M.; Krishnaswamy, S.; Boeglin, M.; Poterszman, A.; Mitschler, A.; Podjarny, A.; Rees, B.; Thierry, J. C.; Moras, D. Science 1991, 252, 1682–1689.
11. Biou, V.; Yaremchuk, A.; Tukalo, M.; Cusack, S. Science 1994, 263, 1404–1410.
12. Cusack, S.; Yaremchuk, A.; Tukalo, M. EMBO J 1996, 15, 2834–2842.
13. Goldgur, Y.; Mosyak, L.; Reshetnikova, L.; Ankilova, V.; Lavrik, O.; Khodyreva, S.; Safro, M. Structure 1996, 5, 59–68.
14. Cusack, S.; Yaremchuk, A.; Krikliviy, I.; Tukalo, M. Structure 1998, 6, 101–108.
15. Cusack, S.; Yaremchuk, A.; Tukalo, M. EMBO J 1996, 15, 6321–6334.
16. Robertus, J. D.; Ladner, J. E.; Finch, J. T.; Rhodes, D.; Brown, R. S.; Clark, B. F.; Klug, A. Nature 1974, 250, 546–51.
17. Kim, S. H.; Suddath, F. L.; Quigley, G. J.; McPherson, A.; Sussman, J. L.; Wang, A. H.; Seeman, N. C.; Rich, A. 1974, Science 185, 435–40.
18. Moras, D.; Comarmond, M. B.; Fischer, J.; Weiss, R.; Thierry, J. C.; Ebel, J. P.; Giege, R. Nature 1980, 288, 669–674.
19. Woo, N. H.; Roe, B. A.; Rich, A. Nature 1980, 286, 346–51.
20. Hayase, Y.; Jahn, M.; Rogers, M. J.; Sylvers, L. A.; Koizumi, M.; Inoue, H.; Ohtsuka, E.; Soll, D. EMBO J 1992, 11, 4159–4165.
21. Jahn, M.; Rogers, M. J.; Soll, D. Nature 1991, 352, 258–260.
22. Putz, J.; Puglisi, J. D.; Florentz, C.; Giege, R. Science 1991, 252, 1696–1699.
23. Frugier, M.; Soll, D.; Giege, R.; Florentz, C. Biochemistry 1994, 33, 9912–9921.
24. Nissen, P.; Kjeldgaard, M.; Thirup, S.; Polekhina, G.; Reshetnikova, L.; Clark, B. F. C.; Nyborg, J. Science 1995, 270, 1464–1472.
25. Witherell, G. W.; Gott, J. M.; Uhlenbeck, O. C. Prog Nucleic Acids Res Mol Biol 1991, 40, 185–220.
26. Valegard, K.; Liljas, L.; Fridborg, K.; Unge, T. Nature 1990, 345, 36–41.
27. Golmohammadi, R.; Valegard, K.; Fridborg, K.; Liljas, L. J Mol Biol 1993, 234, 620–639.
28. Valegard, K.; Murray, J. B.; Stockley, P. G.; Stonehouse, N. J.; Liljas, L. Nature 1994, 371, 623–626.
29. Valegard, K.; Murray, J. B.; Stonehouse, N. J.; Worm, S. v. d.; Stockley, P. G.; Liljas, L. J Mol Biol 1997, 270, 724–738.
30. Borer, P. N.; Lin, Y.; Wang, S.; Roggenbuck, M. W.; Gott, J. M.; Unhlenbeck, O. C.; Pelczer, I. Biochemistry 1995, 34, 6488–6503.

31. van den Worm, S. H. E.; Stonehouse, N. J.; Valegard, K.; Murray, J. B.; Walton, C.; Fridborg, K.; Stockley, P. G.; Liljas, L. Nucleic Acids Res 1998, 26, 1345–1351.
32. Convery, M. A.; Rowsell, S.; Stonehouse, N. J.; Ellington, A. D.; Hirao, I.; Murray, J. B.; Peabody, D. S.; Phillips, S. E. V.; Stockley, P. G. Nature Struct Biol 1998, 5, 133–139.
33. Rowsell, S.; Stonehouse, N. J.; Convery, M. A.; Adams, C. J.; Ellington, A. D.; Hirao, I.; Peabody, D. S.; Stockley, P. G.; Phillips, S. E. Nature Struct Biol 1998, 5, 970–975.
34. Oubridge, C.; Ito, N.; Evans, P. R.; Teo, C. H.; Nagai, K. Nature 1994, 372, 432–438.
35. Jessen, T. H.; Oubridge, C.; Teo, C. H.; Pritchard, C.; Nagai, K. EMBO J 1991, 10, 3447–3456.
36. Allain, F.-H. T.; Gubser, C. C.; Howe, P. W. A.; Nagai, K.; Neuhaus, D.; Varani, G. Nature 1996, 380, 646–650.
37. Howe, P. W. A.; Allain, F. H. T.; Varani, G.; Neuhaus, D. J Biomol NMR 1998, 11, 59–84.
38. Nagai, K.; Oubridge, C.; Jessen, T. H.; Li, J.; Evans, P. R. Nature 1990, 348, 515–520.
39. Gorlach, M.; Burd, C. G.; Dreyfuss, G. J Biol Chem 1994, 269, 23074–23078.
40. Query, C. C.; Bentley, R. C.; Keene, J. D. Cell 1989, 57, 89–101.
41. Price, S. R.; Evans, P. R.; Nagai, K. Nature 1998, 394, 645–650.
42. De Guzman, R. N.; Wu, Z. R.; Stalling, C. C.; Pappalardo, L.; Borer, P. N.; Summers, M. F. Science 1998, 279, 384–388.
43. Morellet, N.; Jullian, N.; De Rocquigny, H.; Maigret, B.; Darlix, J.-L.; Roques, B. P. 1992, EMBO J 11, 3059–3065.
44. Omichinski, J. G.; Clore, G. M.; Sakaguchi, K.; Appella, E.; Gronenborn, A. M. FEBS Lett 1991, 292, 25–30.
45. Summers, M. F.; Henderson, L. E.; Chance, M. R.; Bess, J. W., Jr.; South, T. L.; Blake, P. R.; Sagi, I.; Perez-Alvarado, G.; Sowder, R. C., III; Hare, D. R.; Arthur, L. O. Protein Sci1992, 1, 563–574.
46. Lee, B. M.; De Guzman, R. N.; Turner, B. G.; Tjandra, N.; Summers, M. F. J Mol Biol 1998, 279, 633–649.
47. Pappalardo, L.; Kerwood, D. J.; Pelczer, I.; Borer, P. N. J Mol Biol 1998, 282, 801–818.
48. South, T. L.; Summers, M. F. Protein Sci 1993, 2, 3–19.
49. Varani, G.; Aboul-ela, F.; Allain, F. H. T. Prog Nucleic Magn Reson Spectrosc 1996, 29, 51–127.
50. Ferre-D’Amare, A. R.; Zhou, K.; Doudna, J. A. Nature 1998, 395, 567–574.
51. Jaeger, J.; Restle, T.; Steitz, T. A. EMBO J 1998, 17, 4535–4542.
52. Nicholls, A. GRASP: Graphical Representation and analysis of Surface properties; Columbia University, New York, 1993.