VOLUME 37 NUMBER 5

ACCOUNTS of CHEMICAL RESEARCH®

MAY 2004

Structural Basis of Single-Stranded RNA Recognition

ANA C. MESSIAS AND MICHAEL SATTLER*
Structural and Computational Biology,
European Molecular Biology Laboratory (EMBL),
Meyerhofstrasse 1, 69117 Heidelberg, Germany

Received August 25, 2003

ABSTRACT
RNA is an ancient and highly versatile molecule that plays fundamental roles in all living organisms. Its molecular functions range from being a mediator of genetic information to the regulation of essential cellular processes. These functions are often accomplished in close association with RNA binding proteins. Over the past few years, a considerable number of high-resolution three-dimensional structures of important protein−RNA complexes have been determined. Here, we wish to discuss recent examples and highlight principles and distinct features of single-stranded RNA recognition by conserved RNA binding domains.

Introduction

RNA is presumably one of the earliest molecules in the evolution of life and is thus involved in many different cellular processes.[1] It is a mediator of genetic information and plays essential roles in splicing and translation. In ribozymes, it even exerts catalytic activity.[2] In fact, crystal structures of the ribosome suggest that the catalytic activity for peptide bond formation is exerted by ribosomal RNAs,[3] and it is likely that RNA constitutes the spliceosome catalytic core.[2,4] The importance of RNA has been recently stressed with the discovery of posttranscriptional gene silencing mediated by noncoding RNAs (RNA interference).[5]

Associated with this large variety of RNA molecules is a wide range of RNA binding proteins that process RNA precursors, act as essential cofactors for functional activity, or simply protect RNA from degradation. Not surprisingly, the disruption of protein−RNA interactions can lead to disease (e.g., see ref 6). Therefore, the understanding of the molecular recognition between RNA and proteins and its functional implications is an important subject in structural biology and biomedical research.

The great variety of molecular functions mediated by RNA is reflected in its structural diversity. In contrast to DNA, which preferentially adopts a double-stranded, B-form helical conformation, three-dimensional (3D) structures of RNA comprise single-stranded conformations, double-stranded A-form helices (frequently interrupted by internal loops, mismatches, or bulges), and higher-order tertiary structures. Not surprisingly, RNA binding proteins take advantage of the richness of RNA conformations. Only in the past decade have 3D structures of atomic resolution become available from X-ray and NMR studies. We refer the interested reader to excellent recent reviews describing double-stranded RNA recognition by the double-stranded RNA-binding domain (dsRBD)[7] or by arginine-rich peptides,[8,9] protein−RNA recognition in viral and phage systems,[9] protein−rRNA recognition in the ribosome,[10−12] and protein−tRNA recognition by aminoacyl-tRNA synthetases.[13,14] Here, we wish to highlight recent progress on the structural basis of single-stranded RNA (ssRNA) recognition by conserved RNA binding domains (RRM, KH, and OB-fold), and by modular RNA binding repeats or oligomers (OB-fold, TRAP, Sm proteins, and Pumilio).

The RRM Domain

The RNA recognition motif (RRM, also called RNA binding domain (RBD) or ribonucleoprotein (RNP) domain) is the most abundant and best characterized RNA binding module.[15] Proteins containing RRM domains are implicated in various aspects of the regulation of gene expres-

Ana C. Messias, born in 1970 in Portugal, has a degree in Biochemistry from the University of Lisbon and completed a Ph.D. in Biochemistry at the ITQB, New University of Lisbon. She has been a postdoctoral fellow at EMBL since 2000 (supported by a FCT/FSE fellowship). Her interests are structure determination of biomolecules and the study of the relationship between structure and biological function.

Michael Sattler, born in 1965 in Germany, received a degree in Chemistry from the University of Frankfurt, followed by a Ph.D. with Christian Griesinger. In 1995, he joined Stephen W. Fesik at Abbott Laboratories, Chicago, as a postdoctoral fellow. His research group was established in 1997 and is interested in structural studies of molecular recognition during RNA processing and signal transduction by NMR.

10.1021/ar030034m CCC: $27.50
Published on Web 03/26/2004

sion, such as pre-mRNA splicing by binding to conserved intron RNA sequences and mRNA translation by recognition of the poly(A) tails or of the cap of eukaryotic mRNAs.

The diversity of biological functions associated with RRM proteins is reflected in the wide range of RNA structures and sequences that are recognized with varied affinity and specificity. In some cases, RRM proteins bind the RNA very tightly and with high specificity. For example, the N-terminal RRM domain of the U1A protein binds to approximately seven nucleotides exposed in a U1 snRNA hairpin loop or in the internal loops of its 3′ UTR with a $K_D \approx 10^{-10}$ M, but when the same sequence is presented as ssRNA, the affinity is reduced to $K_D \approx 10^{-6}$ M.[16] In other cases, the binding affinity is lower and the discrimination between different RNAs is poor, suiting proteins that bind their target RNAs transiently. These proteins modulate their affinity and specificity via interactions with additional factors or by using modular RNA binding domains. This is the case for the large (65 kDa) subunit of the heterodimeric U2 auxiliary factor (U2AF^{65}), for which the interactions of the individual domains to a polypyrimidine (Py) ligand are weak ($K_D \approx 10^{-3}$ M)[17] but high affinity ($K_D \approx 10^{-9}$ M) is observed when all three RRMs are present.[18]

The 80−100-residue RRM domain adopts a globular fold comprising a four-stranded antiparallel β-sheet packed against two helices (Figure 1a). The canonical β1-αA-β2-β3-αB-β4 fold can be extended in different ways: by an additional C-terminal helix, αC, as in U1A and U2B″,[19] by a β5 strand as in polypyrimidine tract binding protein (PTB),[20] or by N- and C-terminal extensions as in cleavage stimulation factor (CstF-64).[21] RRM domains are characterized by six- and eight-residue consensus sequences (RNP1 and RNP2), which expose conserved aromatic side chains from the two central strands β3 and β1, respectively. The RRM β-sheet is the primary surface for RNA recognition, while additional contacts mediated by N- and C-terminal residues or loops are important in determining substrate specificity. Interestingly, a growing number of RRMs are found to mediate protein−protein interactions. In most cases, this involves the helical surface of the RRM fold, thus allowing potential RNA binding to the β-sheet platform on the opposite side.[19,22,23] Recently, it was found that even the classical RNA-binding β-sheet surface of an RRM can be employed for protein recognition.[24−26] These observations may provide some clue to the unknown function of other RRM domains, especially in proteins with multiple RRM domains that fail to bind RNA.

The structure of the spliceosomal U1A protein bound to an RNA hairpin[27] represents the first example describing RNA recognition by the RRM domain. This and a number of other RRM−RNA complexes have been reviewed previously.[13,15,28] Here, we wish to highlight two characteristic examples of ssRNA binding by tandem RRMs.

Poly(A)-Binding Protein. The poly(A)-binding protein (PABP) contains four RRM domains and binds to the 3′ poly-A tails of eukaryotic mRNAs in the cytoplasm. This interaction is important for stability of the mRNA but also contributes to a network of molecular interactions that are involved in the initiation of translation. It has been demonstrated that, although single RRM domains do not bind a poly(A) RNA, the two N-terminal RRMs sustain most of the important biological functions of PABP. In the crystal structure of the two N-terminal RRM domains of human PABP bound to a poly(A) ssRNA,[29] the β-sheets of the two RRM domains create an extended RNA-binding platform (Figure 2a). The interdomain linker adopts a helical conformation and is believed to be disordered in the absence of RNA. The poly(A) RNA adopts an extended conformation and runs through the length of the trough

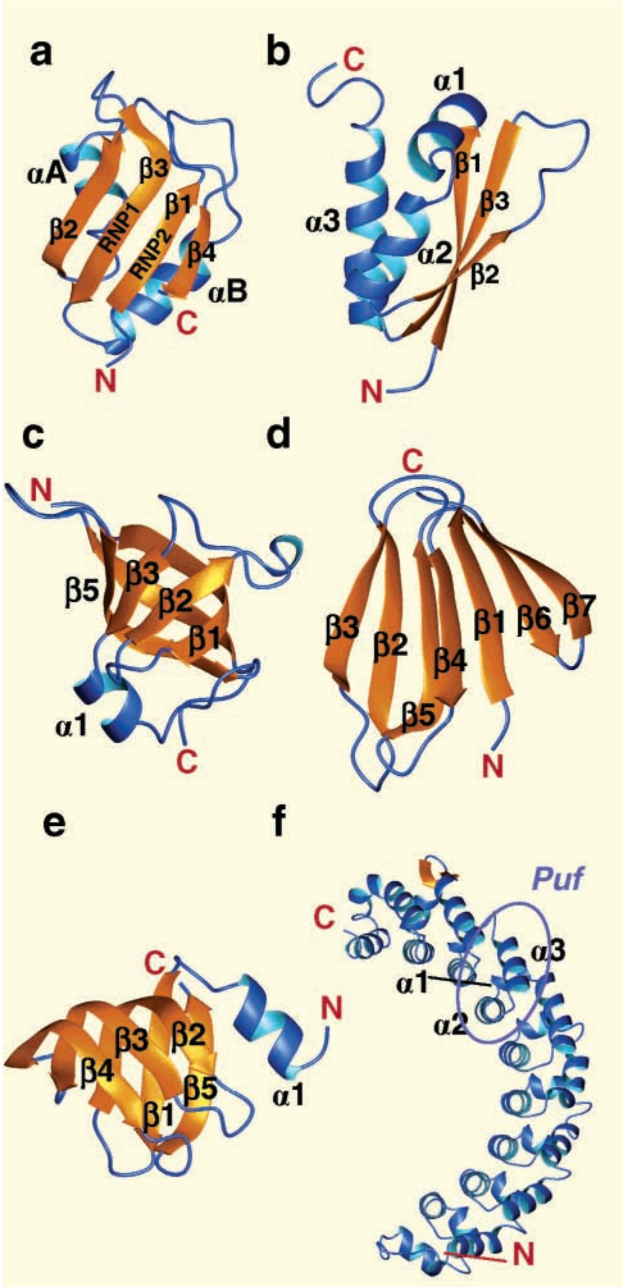

FIGURE 1. Structural folds of ssRNA binding domains: (a) RRM domain (*D. melanogaster* Sxl RRM2, PDB accession code 1B7F);[31] (b) KH domain (human hnRNP K KH3, PDB 1KHM);[44] (c) OB-fold (N-terminal domain of the *E. coli* Rho factor, PDB 1A62);[61] (d) *Bacillus subtilis* TRAP monomer (PDB 1C9S);[52] (e) Sm fold (*P. abyssi* Sm protein, PDB 1M8V);[56] (f) PUM-HD (human Pumilio 1, PDB 1M8X).[59]

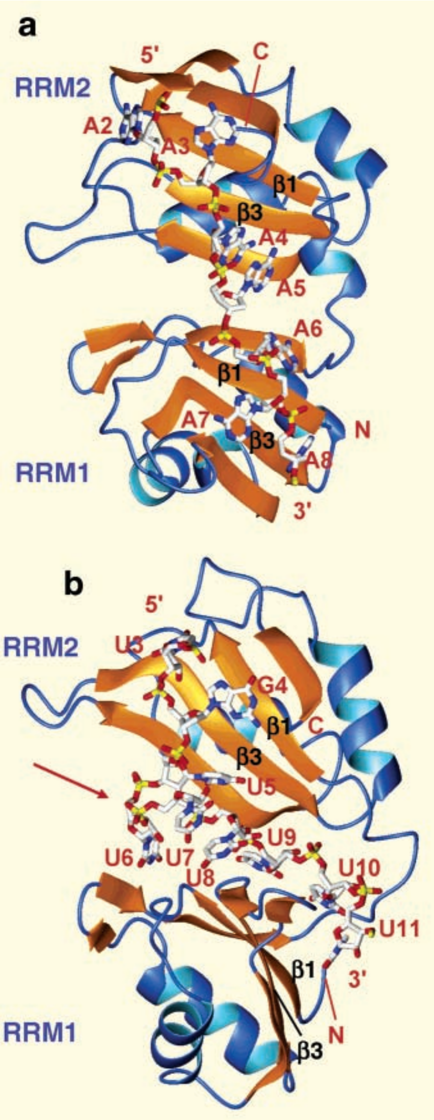

FIGURE 2. Recognition of ssRNA by tandem RRM domains: (a) crystal structure of PABP RRM1−RRM2 complexed with a poly(A) RNA (PDB 1CVJ);[29] (b) crystal structure of the tandem RRMs of Sxl complexed with a *tra* Py-tract ssRNA (PDB 1B7F).[31] The red arrow indicates the kink in the RNA conformation near the V-shaped cleft.

with the 5′ and 3′ halves binding to RRM2 (A2−A5) and RRM1 (A5−A8), respectively. Various combinations of stacking interactions, van der Waals contacts, hydrogen bonds, and salt bridges are employed for adenosine recognition, such that each of the seven nucleotides is recognized in a different way. The most striking aspect of the RNA recognition is the use of extensive intermolecular aromatic−base and intramolecular base−base stacking interactions, favoring a purine base. Sequence specificity is ensured by hydrogen bonding. The phosphate backbone forms hydrogen bonds with tyrosine OH or arginine/lysine amino groups, and in four cases, the ribose 2′-OH interacts with the protein, suggesting that PABP will not bind tightly to poly(A) DNA.

Sex-Lethal Protein. The *Drosophila melanogaster* sex-lethal (Sxl) protein induces female-specific alternative splicing of the *transformer* (*tra*) pre-mRNA. During the sex determination process, Sxl binds tightly to a characteristic uridine-rich Py-tract, preventing binding of the U2AF^{65} to this site and forcing it to bind to the female-specific 3′ splice site. Both RRM domains of Sxl are necessary and sufficient for binding to the *tra* pre-mRNA Py-tract.[30] NMR studies of Sxl RRM1−RRM2 in the absence of RNA indicate that the RRMs do not interact with each other and that the short interdomain linker is flexible.[31] However, in the crystal structure of the tandem RRMs bound to a uridine-rich 12-nucleotide ssRNA (Figure 2b),[31] the β-sheets of the two RRMs do contact each other, forming a V-shaped cleft. The RNA extends along this cleft in an irregular conformation, kinked in the middle and without base pairing. The protein recognizes the sequence 5′-UGUUUUUUU specifically. Like in PABP, the 5′ part of the RNA (UGU) binds to RRM2 and the 3′ region interacts with RRM1. There are several examples of aromatic side chain−RNA base stacking, but there is only one intra-RNA base−base interaction. The RNA backbone is recognized by numerous hydrogen bonds and salt bridges to phosphate groups and to the ribose 2′-OH, consistent with the 10^4-fold weaker binding to a corresponding DNA oligonucleotide.[32] Hydrogen bonds are established with the functional groups of the bases for sequence-specific recognition. In particular, the N3H or O4 groups or both of the eight uridine bases and the 2-amino group of the guanidine base are recognized specifically. Therefore, RNA recognition by Sxl is strictly specific for the 5′-UGUUUUUUU sequence, explaining the reduced affinity of Sxl for Py-tracts with one or more cytidine residues. Since U2AF^{65} prefers cytidine-containing Py-tracts, Sxl prevents the *tra* Py-tract sequence from binding to U2AF,^{65} and the distal female-specific splice site is activated.

The KH Domain

The hnRNP K homology (KH) domain is, alongside the RRM domain, the most abundant nucleic acid-binding domain. The widespread presence of KH domains in eubacteria and eukaryotes suggests that it is an ancient domain. Proteins containing KH domains are involved in the regulation of gene expression at several levels, such as in transcriptional (hnRNP K)[33] or translational (hnRNP K,[34] fragile X mental retardation protein (FMRP)[35]) regulation, splicing (splicing factor 1, SF1)[36] or alternative splicing (Nova-1),[37] and mRNA transport, stability, and localization.[35] As expected from the diverse functional contexts, KH domains bind single-stranded nucleic acids with different selectivities and affinities, ranging from $K_D \approx 10^{-6}$ M for binding to poly(C) ssDNA or ssRNA by the KH3 domain of hnRNP K[38,39] to specific and tight recognition of ssRNA sequences ($K_D \approx 10^{-9}$ M) by the Nova KH3 domain.[40]

A subfamily of KH domains is found in the STAR (signal transduction and activation of RNA) protein family and includes SF1, Sam68 (Src-associated in mitosis; 68 kDa), and others.[41] In STAR proteins, the KH sequence is flanked at the N- and C-termini by two additional homology regions, called QUA1 (quaking homology 1, ~80 residues)

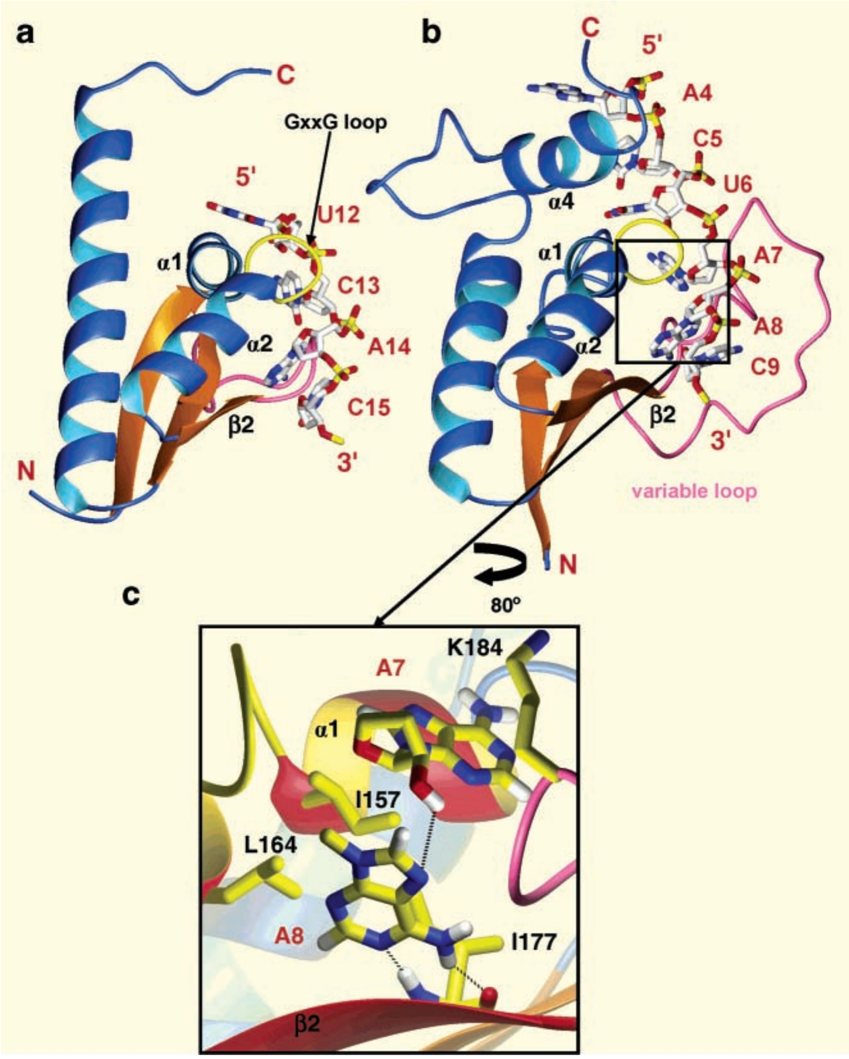

FIGURE 3. RNA recognition by type I KH domains: (a) crystal structure of the Nova-2 KH3 domain complexed with a hairpin loop (PDB 1EC6);[45] (b) solution structure of the SF1 KH-QUA2 domain complexed with a BPS RNA (PDB 1K1G);[36] (c) specific recognition of the branch point adenosine (A8) involving hydrogen bonding to the protein main chain in strand β2.

and QUA2 (~30 residues), respectively, which are also implicated in RNA binding. The QUA1 region is thought to mediate dimerization, thereby potentially increasing RNA binding affinity indirectly,[42] while the QUA2 region of SF1 directly interacts with RNA.[36]

The 70-residue KH domain comprises a three-stranded β-sheet packed against three α-helices. KH domains exhibit two fold variants around a common $\beta\alpha\alpha\beta$ core.[43] The type I KH domain fold (e.g., KH3 of hnRNP K [44]) has a β1-α1-α2-β2-β3-α3 secondary structure (Figure 1b), while the type II fold (e.g., the ribosomal S3 protein[12]) is α1-β1-β2-α2-α3-β3. In both folds, a three-stranded β-sheet platform packs against three helices. However, the β-sheet is antiparallel in type I but mixed in type II. The highly conserved GxxG sequence corresponds to a short loop connecting the two helices of the $\beta\alpha\alpha\beta$ core. A second hallmark of the KH fold is the presence of a loop of variable length connecting β2 and β3 (in type I). Historically, type I and II KH domains have been called maxi-KH and mini-KH, respectively. However, this nomenclature should be avoided since both types are of the same size and are in fact,variations of a related fold.[43] The only two available structures showing ssRNA recognition by type I KH domains[36,45] will be discussed in the following sections.

Nova-2 Protein. The Nova proteins are highly homologous. Nova-1 was found to regulate neuron-specific alternative splicing of two neurotransmitter receptors.[37] Nova proteins contain three KH domains in an asymmetric arrangement with a relatively short interdomain linker between KH1 and KH2 (26−50 residues) and a much larger linker between KH2 and KH3 (179−204 residues).[46] For both proteins, the C-terminal KH domain (KH3) was shown to recognize single-stranded UCAY tetranucleotide sequences in hairpin loops identified by systematic evolution of ligands by exponential enrichment (SELEX) experiments.[40]

In the crystal structure of the Nova-2 KH3 domain bound to a hairpin loop,[45] the bases of the single-stranded U12−C13−A14−C15 loop are splayed out onto a hydrophobic surface comprising helices α1 and α2 and strand β2, leaving the phosphate backbone exposed to the solvent (Figure 3a). The ssRNA binds in a 5′/3′ to C-/N-terminal direction and is gripped between the GxxG loop and the variable loop. Conservation of the glycines in the GxxG loop appears to be required to allow a close approach of the RNA to the protein backbone. A combination of van der Waals contacts to aliphatic side chains and hydrogen bonds involving the ribose 2′-OH, the phosphate backbone, and the functional groups of the bases is used for sequence-specific RNA recognition. Stacking interactions involving aromatic side chains and RNA bases are not used for RNA recognition, and only one intramolecular base−base stacking interaction occurs at the binding interface. Only pyrimidine residues are allowed at the 5′ or 3′ end of the tetranucleotide, but the two central nucleotides must be cytidine and adenine because they establish Watson−Crick-like hydrogen bonds with the protein.

SF1. Initial assembly of the spliceosome involves cooperative binding of SF1 and U2AF to consensus sequences upstream to the 3′ splice site in pre-mRNA introns. In this complex, the KH-QUA2 region of SF1 recognizes the seven-nucleotide branch point sequence (BPS). The solution structure of SF1 KH-QUA2 complexed with an 11-nucleotide ssRNA containing the BPS sequence 5′-UACUAAC has been determined.[36] A surprising structural feature, which is likely to be conserved in other STAR proteins, is an extended KH domain fold where the QUA2 region comprises an additional helix, α4, that packs against the type I KH domain (Figure 3b).

The RNA extends along a hydrophobic surface comprising the QUA2 helix α4 and helices α1 and α2 and strand β2 of the KH domain and is embedded in a groove between the GPRG loop and the long variable loop. The 5′ part of the BPS (A4−C5−U6) is recognized by the C-terminal QUA2 region, while the 3′ part (U6−A7−A8−C9) interacts with the KH domain (Figure 3b). Apart from numerous hydrophobic interactions with aliphatic side chains, the BPS RNA is recognized by hydrogen bonds (including RNA-specific hydrogen bonds involving the ribose 2′-OH) and electrostatic interactions. U6 is recognized at the interface between QUA2 and the KH domain by a combination of hydrogen bonds and van der Waals contacts. The conserved purine A7 is stabilized by a π-cation interaction with a conserved lysine residue in the variable loop (Figure 3c) and contacts a conserved glutamate residue in the β1−α1 loop. The highly conserved branch point adenosine A8 is recognized specifically by Watson−Crick-like hydrogen bonds with the protein and an electrostatic contact with the A7 ribose.

Surprisingly, the recognition of the tetranucleotide U6−A7−A8−C9 by the SF1 KH domain resembles that of the U12−C13−A14−C15 loop by the Nova-2 KH3 domain.[45] Most notably, Watson−Crick-type hydrogen bonds are formed by A8 (SF1) or A14 (Nova2-KH3) to equivalent main chain atoms in strand β2 of the KH domains. However, comparison of the two structures also reveals how the second position is discriminated in different ways. The recognition of C13 by Nova-2 involves hydrogen bonding of the base with Nova-specific glutamate and arginine side chains in helix α1 and the variable loop, respectively, while the corresponding A7 in the BPS is recognized by stacking and electrostatic interactions with amino acids that are conserved only in STAR proteins.

The OB-Fold Domain

The OB-fold is a 70−150 residue domain that was originally named for its oligonucleotide/oligosaccharide binding properties.[47] The diversity of nucleic acid targets and their binding specificity is large, ranging from nonspecific ssDNA binding to sequence-specific recognition of single strands or tertiary folds of DNA or RNA. The *Escherichia coli* transcription terminator factor Rho is an RNA−DNA helicase. It assembles into a hexameric ring with three high-affinity ssRNA/ssDNA binding sites and three low-affinity ssRNA binding sites with preference for poly(C) nucleotides.[48] Nucleic acid binding by Rho is mediated by an OB-fold in its N-terminal domain. The OB-fold comprises a five-stranded β-barrel, β1-β2-β3-α4-β4-β5, arranged in a Greek key motif (Figure 1c). In the cocrystal structure of the Rho N-terminal domain with poly(C) RNA[49] (Figure 4a), two cytidine residues bind specifically to β2 and β3 on the β-sheet surface of the OB-fold, gripped by the β1−β2 loop and helix α4 on one side and the β2−β3 loop on the other. The RNA extends with a 5′ to 3′ polarity from the β3 to the β2 strand. The recognition of the RNA bases is specific for cytosine. The first cytosine is enclosed in a hydrophobic cleft, and the base of the second cytidine stacks with an aromatic side chain of the protein. The functional groups of both bases are engaged in hydrogen bonds. No contacts between the protein and the ribose 2′-OH group are observed, consistent with the observation that this domain binds both ssRNA and ssDNA. Interestingly, the poly(C) recognition by the OB-fold domain of Rho closely resembles the recognition of the anticodon nucleotides in tRNA$^{\mathrm{ASP}}$ by the OB-fold domain of aspartyl-tRNA synthetase.[50] However, in the latter, a GUC trinucleotide is recognized specifically and specific contacts to the 2′-OH groups are observed.

TRAP

The tryptophan RNA binding attenuation protein (TRAP) is a 70-residue protein that regulates expression of the L-tryptophan biosynthetic genes in several bacilli by binding to the 5′ noncoding leader region of the operon mRNA. TRAP binds to this RNA when activated by bound L-tryptophan to form a "terminator" loop, which leads to transcription termination by preventing the formation of an "antiterminator" stem−loop structure.[51]

The TRAP protein adopts an antiparallel β-sandwich fold (Figure 1d) with the tryptophan inserted between the two β-platforms. TRAP assembles into an 11-mer symmetric ring (Figure 4b), both in the absence and in the presence of L-tryptophan. In the crystal structure of a complex of TRAP and a 53-nucleotide ssRNA, the 11 GAGAU repeats have essentially the same 3D structure and bind to the outer surface of the 80 Å diameter ring.[52] The proper curvature for binding to the oligomeric TRAP is achieved by a right-handed A-helical conformation of the central AGA nucleotides and a compensating left-

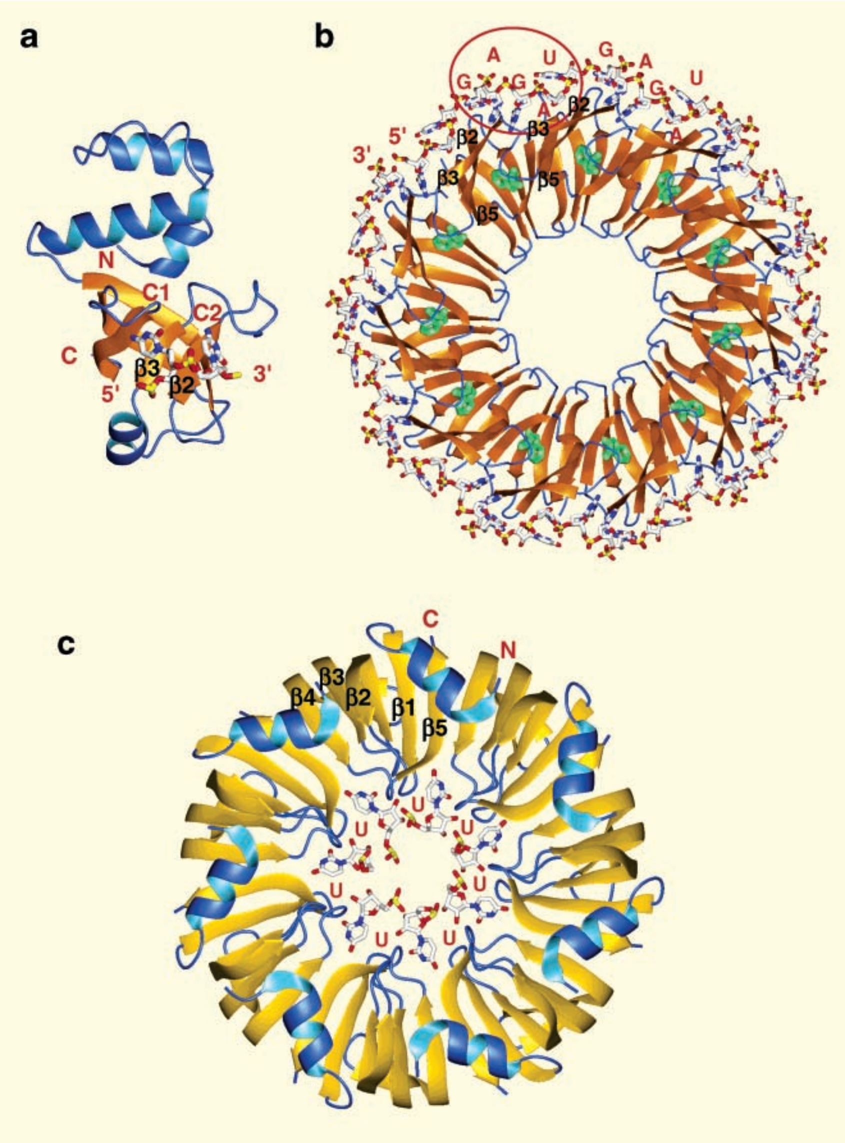

FIGURE 4. Modular RNA recognition by RNA binding repeats and oligomeric RNA binding domains: (a) crystal structure of one monomer of the N-terminal *E. coli* Rho domain complexed with a poly(C) ssRNA (PDB 2A8V);[49] (b) crystal structure of the TRAP 11-mer ring complexed with a ssRNA containing 11 GAGAU repeats (PDB 1C9S)[52] (the L-tryptophan amino acid inserted in the middle of the β-sandwich is shown in green; one of the GAGAU repeats is circled in red); (c) crystal structure of the Sm core from *P. abyssi* bound to a poly(U) ssRNA (PDB 1M8V).[56]

handed turn-like conformation of the other two nucleotides. The 11 GAG triplets are inserted between two consecutive TRAP molecules and interact with the protein, while the bases of the AU dinucleotide spacer stack with each other. Hydrogen bonds lead to highly specific recognition of A2 and G3 in the GAG triplets. No interactions between the protein and the phosphate backbone are observed, and only one hydrogen bond is established with the ribose 2′-OH of G3 in five of the 11 units. These hydrogen bonds seem to be critical in distinguishing between RNA and DNA analogues, as an RNA ligand with a deoxyribose at this position binds to TRAP with 10^4-fold reduced affinity.[53]

Sm Proteins

The seven Sm proteins (B/B′, D_3, D_2, D_1, E, F, and G) are shared by the uridine-rich small nuclear U snRNPs involved in pre-mRNA splicing, forming a common core of the snRNP complexes. These proteins bind to the uridine-rich Sm-site sequence 5′-AAUUUUUGA in the snRNAs. The Sm proteins exhibit a common ~80-residue fold that consists of an N-terminal α-helix followed by a strongly bent five-stranded antiparallel β-sheet (Figure 1e). In the presence of Sm-site RNA, the seven Sm proteins are thought to assemble into a oligomeric ring-like structure with the poly(U) ligand binding along its inner surface.[54] This heptameric ring, with an outer diameter of ~70 Å and a central hole of ~20 Å diameter, fits well into the electron density envelope of the complete U1 snRNP determined by electron microscopy.[55]

In the crystal structure of the archaeal *Pyrococcus abyssi* Sm-related protein complexed with a poly(U) ssRNA ligand[56] (Figure 4c), each uridine is recognized by three conserved residues, which are located in the $\beta2-\beta3$ and

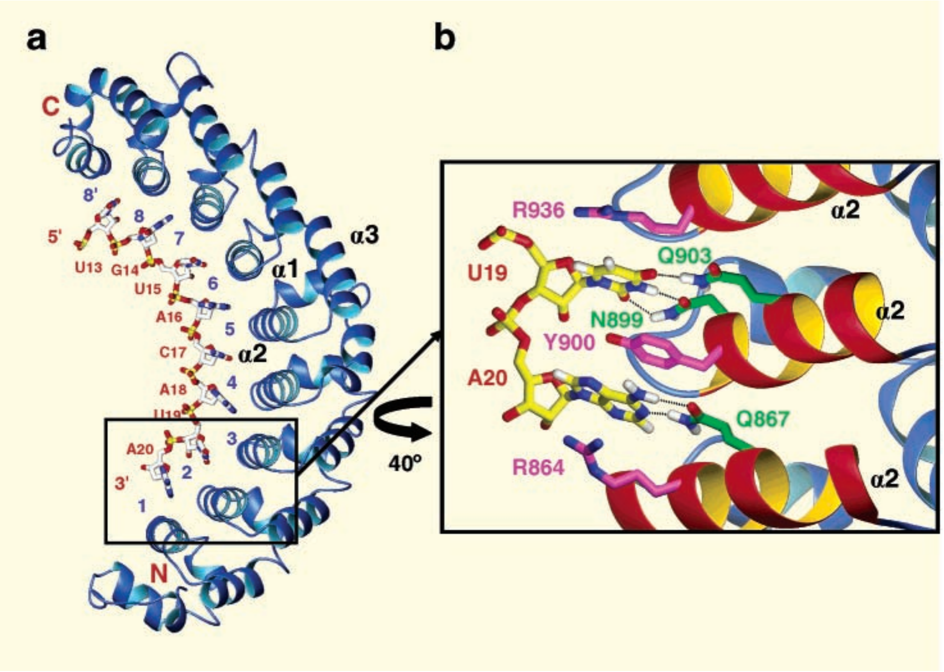

FIGURE 5. Crystal structure (a) of the human Pumilio1 PUM-HD complexed with a 10-nucleotide ssRNA (PDB 1M8Y)[59] and (b) RNA recognition by *Puf* repeats.

β4−β5 loops of a given Sm monomer. The uridine bases are stacked between an aromatic and a basic residue, and the Watson−Crick functional groups are recognized specifically by hydrogen bonding to an asparagine side chain. Interestingly, the base recognition resembles the RNA recognition by Pumilio, even though the protein folds are dissimilar (vide infra). Based on the archaeal Sm/RNA complex, a refined model for the human Sm core and its interaction with Sm site RNA was proposed, which, however, still awaits experimental verification.

Pumilio-Homology Domain

The Pumilio-homology domain (PUM-HD) proteins (or PUF, named after its two founding members *Drosophila* Pumilio and *Caenorhabditis elegans* fem-3 mRNA binding factor) are a family of structurally related proteins that bind to 3′ untranslated regions and modulate mRNA expression in a wide variety of eukaryotic species.[57] In combination with other regulatory proteins, they are involved in the control of diverse developmental events by enhancing turnover or repressing translation. The PUM-HD contains eight consecutive so-called *Puf* repeats, flanked N- and C-terminally by two *Puf*-related sequences. Each ∼40-residue repeat forms a three-helical bundle with a short helix, α1, and two long helices, α2 and α3. The repeats line up to form a curved structure, defining approximately half a ring of 80 Å diameter (Figure 1f).[57] A “core consensus” of hydrophobic, basic, and acidic residues in helix α2 is located at the concave surface of the protein. Notably, the entire PUM-HD is necessary and sufficient for RNA binding and protein interaction.[58]

In the crystal structure of the human Pumilio1 PUM-HD complexed with a 10-nucleotide sequence, 5′-AUUGUACAUA, the ssRNA extends along the concave protein surface in a 5′/3′ to C-/N-terminal direction (Figure 5a).[59] Few intramolecular contacts are observed for the RNA. The bases contact the protein surface, while the phosphate groups are exposed to the solvent. The eight nucleotides 5′-UGUACAUA are individually recognized by three conserved amino acids in the *Puf* repeats 8−1, respectively (Figure 5b). The bases are stacked between the side chains of amino acids at equivalent positions (residue 13) in helix α2 of the repeats. Interestingly, these stacking interactions involve alternating aromatic and basic side chains. Specific base recognition is achieved by hydrogen bonding with residues at position 12 (asparagine, cysteine, and serine) and 16 (glutamine and glutamate) in helix α2. Even though the RNA bases mediate the primary contacts with the protein, recognition of the RNA backbone is important, because PUM-HD binds RNA more than 2500-fold more strongly than a corresponding DNA sequence. In fact, the 2′-OH groups establish intramolecular hydrogen bonds to the phosphate oxygen atoms and water-mediated intermolecular hydrogen bonds to the protein.

Principles of ssRNA Recognition

The growing database of 3D structures of protein−ssRNA complexes provides insight into the principles of ssRNA recognition. As a common feature, 2−10 nucleotides per domain or monomer are bound in an extended conformation along the protein binding surface with few intra-RNA interactions and usually in a 5′/3′ to C-/N-terminal direction. The bases are readily accessible in an extended ssRNA conformation and are recognized specifically by a combination of hydrophobic/stacking interactions and hydrogen bonding to the protein.

The versatility employed by these RNA binding domains is remarkable, as demonstrated by the seven different modes employed for adenosine recognition by PABP or the ability of a given ssRNA binding domain (e.g., RRMs in PABP and Sxl) to bind specifically to poly(A) or poly(U) sequences. Interestingly, similar modes of RNA recognition are employed by diverse structural elements, ranging from β-sheet surfaces (RRM) to entirely α-helical folds (PUM-HD) or loops (Sm proteins). This shows that specific ssRNA recognition can be achieved by a great variety of protein folds.

Nevertheless, some general rules are discernible. For example, a hallmark of RNA recognition by RRM domains is stacking of exposed aromatic side chains in the RNP1 and RNP2 sequence motifs with RNA bases. In contrast, the hydrophobic RNA binding surface in KH domains is lined with aliphatic side chains, and aromatic/base stacking is not observed, consistent with the paucity of aromatic residues in KH domains. Another unique and likely conserved feature is hydrogen bonding of the Watson−Crick functional groups of the third base in the tetranucleotide ligand to the peptide main chain in strand $\beta2$ of the KH domain.

The selectivity of RNA binding domains for RNA versus DNA ligands is related to the extent of protein−RNA contacts involving the ribose 2′-OH group and varies considerably. Some domains (PUM-HD, Sxl) are strictly specific for RNA, while others (Rho OB-fold, hnRNP K KH domain) are only poorly selective.

The conformational flexibility of ssRNA usually leads to an induced fit of the RNA to the protein surface. In some cases (e.g., Nova-2 and U1A), the cognate RNA sequence is within the single-stranded loop of a hairpin structure, thereby presumably reducing the loss of conformational entropy associated with binding. However, induced fit of the protein is also observed and can play important functional roles in the regulation of additional molecular interactions. For example, upon RNA binding, helix αC in the U1A RRM reorients[60] and helix αC in the CstF-64 RRM unfolds.[21]

An interesting aspect of many RNA binding proteins is their modular arrangement. Frequently, RRMs are required in tandem to form a high-affinity binding platform, as seen for PABP and Sxl, where the two domains interact and the interdomain linker is actively involved in RNA recognition. RNA binding by multiple RRMs is often associated with a change in interdomain orientation and mobility. These conformational changes can play important roles in the regulation of biological activity, and the resulting protein−RNA complexes may present a binding surface for additional factors, forming multimeric regulatory complexes.

ssRNA recognition by proteins involving multiple repeats or oligomerization of a structural fold, such as by Rho, Pumilio, TRAP, or Sm, is modular and less complex than independent RNA binding domains − single nucleotides or short sequence stretches are recognized by individual RNA binding repeats or monomers via redundant interactions. Nevertheless, the context of multimeric repeats or the oligomeric state is essential, since additional RNA interactions are mediated by the interface between individual repeats or monomers.

Concluding Remarks

ssRNA binding domains recognize RNA sequences with a wide range of affinities and specificities. This versatility correlates with the various roles played by RNA binding proteins in the control of gene expression. While some principles can be deduced, more structural data are required to improve our understanding of protein−RNA recognition and to start deciphering a recognition code that, in the future, might allow prediction of the RNA binding properties of a protein from its structure.

References

(1) *The RNA world*, 2nd ed.; Gesteland, R. F., Cech, T. R., Atkins, J. F., Eds.; Cold Spring Harbor Laboratory Press: New York, 1999.
(2) Cech, T. R. Ribozymes, the first 20 years. *Biochem. Soc. Trans.* **2002**, *30*, 1162−1166.
(3) Steitz, T. A.; Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **2003**, *28*, 411−418.
(4) Collins, C. A.; Guthrie, C. The question remains: is the spliceosome a ribozyme? *Nat. Struct. Biol.* **2000**, *7*, 850−854.
(5) Denli, A. M.; Hannon, G. J. RNAi: an ever-growing puzzle. *Trends Biochem. Sci.* **2003**, *28*, 196−201.
(6) Faustino, N. A.; Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **2003**, *17*, 419−437.
(7) Fierro-Monti, I.; Mathews, M. B. Proteins binding to duplexed RNA: one motif, multiple functions. *Trends Biochem. Sci.* **2000**, *25*, 241−246.
(8) Weiss, M. A.; Narayana, N. RNA recognition by arginine-rich peptide motifs. *Biopolymers* **1998**, *48*, 167−180.
(9) Patel, D. J. Adaptive recognition in RNA complexes with peptides and protein modules. *Curr. Opin. Struct. Biol.* **1999**, *9*, 74−87.
(10) Brodersen, D. E.; Clemons, W. M., Jr.; Carter, A. P.; Wimberly, B. T.; Ramakrishnan, V.; Morgan-Warren, R. J.; Vonrhein, C.; Hartsch, T. Crystal structure of the 30 S ribosomal subunit from *Thermus thermophilus*: structure of the proteins and their interactions with 16 S RNA. *J. Mol. Biol.* **2002**, *316*, 725−768.
(11) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **2000**, *289*, 905−920.
(12) Wimberly, B. T.; Brodersen, D. E.; Clemons, W. M., Jr.; Morgan-Warren, R. J.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. Structure of the 30S ribosomal subunit. *Nature* **2000**, *407*, 327−339.
(13) De Guzman, R. N.; Turner, R. B.; Summers, M. F. Protein-RNA recognition. *Biopolymers* **1998**, *48*, 181−195.
(14) Cusack, S. RNA-protein complexes. *Curr. Opin. Struct. Biol.* **1999**, *9*, 66−73.
(15) Varani, G.; Nagai, K. RNA recognition by RNP proteins during RNA processing. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 407−445.
(16) Hall, K. B. Interaction of RNA hairpins with the human U1A N-terminal RNA binding domain. *Biochemistry* **1994**, *33*, 10076−10088.
(17) Ito, T.; Muto, Y.; Green, M. R.; Yokoyama, S. Solution structures of the first and second RNA-binding domains of human U2 small nuclear ribonucleoprotein particle auxiliary factor (U2AF(65)). *EMBO J.* **1999**, *18*, 4523−4534.
(18) Zamore, P. D.; Patton, J. G.; Green, M. R. Cloning and domain structure of the mammalian splicing factor U2AF. *Nature* **1992**, *355*, 609−614.
(19) Price, S. R.; Evans, P. R.; Nagai, K. Crystal structure of the spliceosomal U2B″-U2A′ protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **1998**, *394*, 645−650.
(20) Conte, M. R.; Grune, T.; Ghuman, J.; Kelly, G.; Ladas, A.; Matthews, S.; Curry, S. Structure of tandem RNA recognition motifs from polypyrimidine tract binding protein reveals novel features of the RRM fold. *EMBO J.* **2000**, *19*, 3132−3141.
(21) Perez Canadillas, J. M.; Varani, G. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J.* **2003**, *22*, 2821−2830.

(22) Kielkopf, C. L.; Rodionova, N. A.; Green, M. R.; Burley, S. K. A novel peptide recognition mode revealed by the X-ray structure of a core U2AF^{35}/U2AF^{65} heterodimer. *Cell* **2001**, *106*, 595–605.

(23) Selenko, P.; Gregorovic, G.; Sprangers, R.; Stier, G.; Rhani, Z.; Kramer, A.; Sattler, M. Structural basis for the molecular recognition between human splicing factors U2AF^{65} and SF1/mBBP. *Mol. Cell* **2003**, *11*, 965–976.

(24) Fribourg, S.; Gatfield, D.; Izaurralde, E.; Conti, E. A novel mode of RBD-protein recognition in the Y14-Mago complex. *Nat. Struct. Biol.* **2003**, *10*, 433–439.

(25) Shi, H.; Xu, R. M. Crystal structure of the *Drosophila* Mago nashi-Y14 complex. *Genes Dev.* **2003**, *17*, 971–976.

(26) Lau, C. K.; Diem, M. D.; Dreyfuss, G.; Van Duyne, G. D. Structure of the Y14-Magoh core of the exon junction complex. *Curr. Biol.* **2003**, *13*, 933–941.

(27) Oubridge, C.; Ito, N.; Evans, P. R.; Teo, C. H.; Nagai, K. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **1994**, *372*, 432–438.

(28) Perez-Canadillas, J. M.; Varani, G. Recent advances in RNA-protein recognition. *Curr. Opin. Struct. Biol.* **2001**, *11*, 53–58.

(29) Deo, R. C.; Bonanno, J. B.; Sonenberg, N.; Burley, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **1999**, *98*, 835–845.

(30) Samuels, M.; Deshpande, G.; Schedl, P. Activities of the Sex-lethal protein in RNA binding and protein:protein interactions. *Nucleic Acids Res.* **1998**, *26*, 2625–2637.

(31) Handa, N.; Nureki, O.; Kurimoto, K.; Kim, I.; Sakamoto, H.; Shimura, Y.; Muto, Y.; Yokoyama, S. Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature* **1999**, *398*, 579–585.

(32) Kanaar, R.; Lee, A. L.; Rudner, D. Z.; Wemmer, D. E.; Rio, D. C. Interaction of the sex-lethal RNA binding domains with RNA. *EMBO J.* **1995**, *14*, 4530–4539.

(33) Krecic, A. M.; Swanson, M. S. hnRNP complexes: composition, structure, and function. *Curr. Opin. Cell Biol.* **1999**, *11*, 363–371.

(34) Ostareck-Lederer, A.; Ostareck, D. H.; Hentze, M. W. Cytoplasmic regulatory functions of the KH-domain proteins hnRNPs K and E1/E2. *Trends Biochem. Sci.* **1998**, *23*, 409–411.

(35) Perrone-Bizzozero, N.; Bolognani, F. Role of HuD and other RNA-binding proteins in neural development and plasticity. *J. Neurosci. Res.* **2002**, *68*, 121–126.

(36) Liu, Z.; Luyten, I.; Bottomley, M. J.; Messias, A. C.; Houngninou-Molango, S.; Sprangers, R.; Zanier, K.; Kramer, A.; Sattler, M. Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **2001**, *294*, 1098–1102.

(37) Jensen, K. B.; Dredge, B. K.; Stefani, G.; Zhong, R.; Buckanovich, R. J.; Okano, H. J.; Yang, Y. Y.; Darnell, R. B. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* **2000**, *25*, 359–371.

(38) Braddock, D. T.; Baber, J. L.; Levens, D.; Clore, G. M. Molecular basis of sequence-specific single-stranded DNA recognition by KH domains: solution structure of a complex between hnRNP K KH3 and single-stranded DNA. *EMBO J.* **2002**, *21*, 3476–3485.

(39) Messias, A. C.; Backe, P.; Cusack, S.; Sattler, M. Unpublished results.

(40) Jensen, K. B.; Musunuru, K.; Lewis, H. A.; Burley, S. K.; Darnell, R. B. The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5740–5745.

(41) Vernet, C.; Artzt, K. STAR, a gene family involved in signal transduction and activation of RNA. *Trends Genet.* **1997**, *13*, 479–484.

(42) Chen, T.; Richard, S. Structure–function analysis of Qk1: a lethal point mutation in mouse quaking prevents homodimerization. *Mol. Cell Biol.* **1998**, *18*, 4863–4871.

(43) Grishin, N. V. KH domain: one motif, two folds. *Nucleic Acids Res.* **2001**, *29*, 638–643.

(44) Baber, J. L.; Libutti, D.; Levens, D.; Tjandra, N. High precision solution structure of the C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K, a c-myc transcription factor. *J. Mol. Biol.* **1999**, *289*, 949–962.

(45) Lewis, H. A.; Musunuru, K.; Jensen, K. B.; Edo, C.; Chen, H.; Darnell, R. B.; Burley, S. K. Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* **2000**, *100*, 323–332.

(46) Lewis, H. A.; Chen, H.; Edo, C.; Buckanovich, R. J.; Yang, Y. Y.; Musunuru, K.; Zhong, R.; Darnell, R. B.; Burley, S. K. Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure* **1999**, *7*, 191–203.

(47) Theobald, D. L.; Mitton-Fry, R. M.; Wuttke, D. S. Nucleic Acid Recognition by OB-Fold Proteins. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 115–133.

(48) Skordalakes, E.; Berger, J. M. Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell* **2003**, *114*, 135–146.

(49) Bogden, C. E.; Fass, D.; Bergman, N.; Nichols, M. D.; Berger, J. M. The structural basis for terminator recognition by the Rho transcription termination factor. *Mol. Cell* **1999**, *3*, 487–493.

(50) Cavarelli, J.; Rees, B.; Ruff, M.; Thierry, J. C.; Moras, D. Yeast tRNA(Asp) recognition by its cognate class II aminoacyl-tRNA synthetase. *Nature* **1993**, *362*, 181–184.

(51) Babitzke, P. Regulation of tryptophan biosynthesis: Trp-ing the TRAP or how *Bacillus subtilis* reinvented the wheel. *Mol. Microbiol.* **1997**, *26*, 1–9.

(52) Antson, A. A.; Dodson, E. J.; Dodson, G.; Greaves, R. B.; Chen, X.; Gollnick, P. Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* **1999**, *401*, 235–242.

(53) Elliott, M. B.; Gottlieb, P. A.; Gollnick, P. Probing the TRAP–RNA interaction with nucleoside analogs. *RNA* **1999**, *5*, 1277–1289.

(54) Kambach, C.; Walke, S.; Young, R.; Avis, J. M.; de la Fortelle, E.; Raker, V. A.; Lührmann, R.; Li, J.; Nagai, K. Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **1999**, *96*, 375–387.

(55) Stark, H.; Dube, P.; Lührmann, R.; Kastner, B. Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. *Nature* **2001**, *409*, 539–542.

(56) Thore, S.; Mayer, C.; Sauter, C.; Weeks, S.; Suck, D. Crystal Structures of the *Pyrococcus abyssi* Sm Core and Its Complex with RNA. COMMON FEATURES OF RNA BINDING IN ARCHAEA AND EUKARYA. *J. Biol. Chem.* **2003**, *278*, 1239–1247.

(57) Wickens, M.; Bernstein, D. S.; Kimble, J.; Parker, R. A PUF family portrait: 3′UTR regulation as a way of life. *Trends Genet.* **2002**, *18*, 150–157.

(58) Zhang, B.; Gallegos, M.; Puoti, A.; Durkin, E.; Fields, S.; Kimble, J.; Wickens, M. P. A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature* **1997**, *390*, 477–484.

(59) Wang, X.; McLachlan, J.; Zamore, P. D.; Hall, T. M. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **2002**, *110*, 501–512.

(60) Varani, L.; Gunderson, S. I.; Mattaj, I. W.; Kay, L. E.; Neuhaus, D.; Varani, G. The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. *Nat. Struct. Biol.* **2000**, *7*, 329–335.

(61) Allison, T. J.; Wood, T. C.; Briercheck, D. M.; Rastinejad, F.; Richardson, J. P.; Rule, G. S. Crystal structure of the RNA-binding domain from transcription termination factor rho. *Nat. Struct. Biol.* **1998**, *5*, 352–356.

AR030034M