RNA-binding proteins: modular design for efficient function

Bradley M. Lunde[‡§], *Claire Moore*[¶] *and Gabriele Varani*[*‡]

Abstract | Many RNA-binding proteins have modular structures and are composed of multiple repeats of just a few basic domains that are arranged in various ways to satisfy their diverse functional requirements. Recent studies have investigated how different modules cooperate in regulating the RNA-binding specificity and the biological activity of these proteins. They have also investigated how multiple modules cooperate with enzymatic domains to regulate the catalytic activity of enzymes that act on RNA. These studies have shown how, for many RNA-binding proteins, multiple modules define the fundamental structural unit that is responsible for biological function.

Ribonucleoprotein
(RNP). A complex that contains proteins and RNA. The RNP motif refers to the two conserved sequence elements found in the RNA-recognition motif (RRM) (in its two central β-strands) that participate in RNA recognition and identify the RRM domain at the sequence level.

*Department of Chemistry and ‡Biochemistry, §Biomolecular Structure and Design, University of Washington, Seattle, Washington 98195, USA. ¶Department of Molecular Biology and Microbiology, Tufts University School of Medicine, 136 Harrison Avenue, Boston, Massachusetts 02111, USA. Correspondence to G.V. e-mail: varani@chem.washington.edu
doi:10.1038/nrm2178
Published online 2 May 2007

RNA is rarely at a loss for companions; as soon as RNA is transcribed, ribonucleoproteins (RNPs) form co-transcriptionally on the nascent transcript and participate in processing, nuclear export, transport and localization[1]. The dynamic association of these proteins with RNA defines the lifetime, cellular localization, processing and the rate at which a specific mRNA is translated.

The diversity of functions of RNA-binding proteins would suggest a correspondingly large diversity in the structures that are responsible for RNA recognition. However, most RNA-binding proteins are built from few RNA-binding modules (TABLE 1). Instead, the large structural diversity of substrates is accommodated by the presence of multiple copies of these RNA-binding domains (RBDs) presented in various structural arrangements to expand the functional repertoire of these proteins[2] (FIG. 1). Modules of the same or of different structural types combine to create versatile macromolecular binding surfaces to define the specificity of these proteins and combine with enzymatic domains to define the targets of the enzymes and regulate catalytic activity (FIG. 2). To understand the function of RNA-binding proteins, it is therefore important to know how these domains function together as RNA-recognition units.

Here, we focus on how RNA-binding modules are combined and arranged to facilitate a myriad of different interactions and regulatory events. We first illustrate general themes as to how modularity facilitates function. We then briefly summarize the principles of RNA recognition by individual RBDs as a necessary prologue to the subsequent discussion of how specific combinations of modules cooperate functionally and structurally. The reader is referred to several excellent reviews that discuss in greater detail the molecular mechanisms that are used by individual domains to recognize specific RNAs[3–6].

Modularity facilitates function

Many cellular processes, including those involved in intracellular signalling and the extracellular matrix[7–9], rely on proteins that are constructed through multiple repeats of a few basic modular units. The advantages to constructing a protein with a modular architecture arise from the resulting versatility. By existing in multiple copies (FIG. 1), these modules endow a protein with the ability to bind RNA with higher specificity and affinity than would be possible with individual domains, which often bind short RNA stretches with weak affinity. Therefore, by constructing an interaction surface through multiple modules, high affinity and specificity for a particular target can be obtained by combining multiple weak interactions. These weak interactions make it easier to regulate the formation of these complexes by disassembling them when needed. Furthermore, these multiple binding sites can evolve independently. The modular architecture is also ideally suited to construct proteins that match in their RNA specificity the poorly conserved sequence features that are observed in splicing and 3′-end processing sites of eukaryotic mRNAs[10–12].

The first effect of providing a protein with multiple domains is therefore that the protein itself can recognize a much longer stretch of nucleic acids than would be possible with a single domain (FIG. 2a, left). This modularity also allows proteins to recognize sequences that are either separated by an intervening stretch of nucleotides (FIG. 2a, centre) or that belong to different RNAs (FIG. 2a, right).

Table 1 | **Common RNA-binding domains and their properties**

Domain	Topology	RNA-recognition surface	Protein–RNA interactions	Representative structures (PDB ID)
RRM	αβ	Surface of β-sheet	Interacts with about four nucleotides of ssRNA through stacking, electrostatics and hydrogen bonding	U1A N-terminal RRM[18] (1URN)
KH (type I and type II)	αβ	Hydrophobic cleft formed by variable loop between β2, β3 and GXXG loop. Type II: same as type I, except variable loop is between α2 and β2	Recognizes about four nucleotides of ssRNA through hydrophobic interactions between non-aromatic residues and the bases; sugar-phosphate backbone contacts from the GXXG loop, and hydrogen bonding to bases	Nova-1 KH3 (type I)[41] (1EC6), NusA (type II)[37] (2ASB)
dsRBD	αβ	Helix α1, N-terminal portion of helix α2, and loop between β1 and β2	Shape-specific recognition of the minor–major–minor groove pattern of dsRNA through contacts to the sugar-phosphate backbone; specific contacts from the N-terminal α-helix to RNA in some proteins	dsRBD3 from Staufen[51] (1EKZ)
ZnF-CCHH	αβ	Primarily residues in α-helices	Protein side chain contacts to bulged bases in loops and through electrostatic interactions between side chains and the RNA backbone	Fingers 4–6 of TFIIIA[56] (1UN6)
ZnF-CCCH	Little regular secondary structure	Aromatic side chains form hydrophobic binding pockets for bases that make direct hydrogen bonds to protein backbone	Stacking interactions between aromatic residues and bases create a kink in RNA that allows for the direct recognition of Watson–Crick edges of the bases by the protein backbone	Fingers 1 and 2 of TIS11d[57] (1RGO)
S1	β	Core formed by two β-strands with contributions from surrounding loops	Stacking interactions between bases and aromatic residues and hydrogen bonding to the bases	Ribonuclease II[121] (2IX1), exosome[99] (2NN6)
PAZ	αβ	Hydrophobic pocket formed by OB-like β-barrel and small αβ motif	Recognizes single-stranded 3′ overhangs of siRNA through stacking interactions and hydrogen bonds	PAZ[73] (1SI3), Argonaute[76] (1U04), Dicer[72] (2FFL)
PIWI	αβ	Highly conserved pocket, including a metal ion that is bound to the exposed C-terminal carboxylate	Recognizes the defining 5′ phosphate group in the siRNA guide strand with a highly conserved binding pocket that includes a metal ion	PIWI[75] (1YTU), Argonaute (1U04)[76]
TRAP	β	Edges of β-sheets between each of the 11 subunits that form the entire protein structure	Recognizes the GAG triplet through stacking interactions and hydrogen bonding to bases; limited contacts to the backbone	TRAP[122] (1C9S)
Pumilio	α	Two repeats combine to form binding pocket for individual bases; helix α2 provides specificity-determining residues	Binding pockets for bases provided by stacking interactions; specificity dictated by hydrogen bonds to the Watson–Crick face of a base by two amino acids in helix α2	Pumilio[84] (1M8Y)
SAM	α	Hydrophobic cavity between three helices surrounded by an electropositive region	Shape-dependent recognition of RNA stem–loop, mainly through interactions with the sugar-phosphate backbone and a single base in the loop	Vts1[123] (2ESE)

dsRBD, double-stranded RNA-binding domain; KH, K-homology; OB-like, oligonucleotide/oligosaccharide binding-like; PDB ID, Protein Data Bank identification; RRM, RNA-recognition motif; siRNA, small interfering RNA; ssRNA, single-stranded RNA; ZnF, zinc finger.

The specificity of individual domains in a protein is functionally important, but so is the way in which domains are arranged relative to each other. This is reflected in evolution: higher levels of conservation are often found between domains that occupy the same position in orthologous proteins, as opposed to domains in the same protein but in a different position. For example, in both the splicing factor U2 auxiliary factor (U2AF) subunit 65 and in the poly(A)-binding protein (PABP), the RNA-recognition motif-1 (RRM1) in yeast is more similar to the RRM1 of the human protein than it is to the RRM3 or RRM4 of the yeast protein.

Much of the ability of these proteins to recognize RNA specifically depends on the linker between the two domains. Long linkers are generally disordered and allow the two domains to recognize a diverse set of targets, as shown in the centre and right panels of FIG. 2a, whereas short linkers predispose the domains to bind to a contiguous stretch of nucleic acids (FIG. 2a, left side). When this occurs, the linker domain generally becomes ordered, forming a short α-helix in response to RNA binding that positions the two domains relative to one another and sometimes contacts RNA directly[13–16]. In these situations, interdomain sequences are as well conserved as, or better conserved than, the domains themselves[17] because the precise positioning of domains facilitates their function.

The modular architecture allows a protein to topologically arrange the generally flexible RNA for a particular function (FIG. 2b). Conversely, the proteins themselves can be topologically organized to interact with a particular RNA structure (FIG. 2c); for example, additional domains can be used (FIG. 2c, yellow oval) to organize the RBDs.

Last, the combination of enzymatic domains and RBDs provides ways to regulate catalytic activity. In FIG. 2d, we outline a situation in which the active site of an enzyme is occluded by the presence of an RBD. In the presence of the substrate RNA, the RBD binds its target, thereby releasing the enzyme from its inactive state.

Orthologous proteins
Proteins that are direct evolutionary counterparts, that retain the same function in different organisms and that have arisen due to speciation events but not through the process of gene duplication (paralogous proteins).

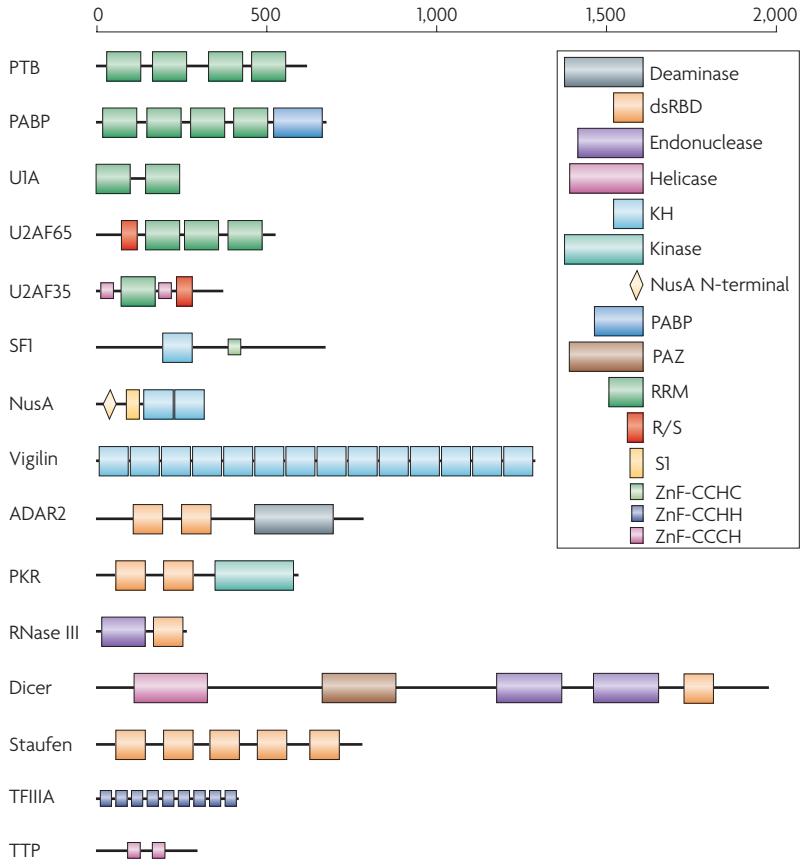

Figure 1 | **Many RNA-binding proteins have a modular structure.** Representative examples from some of the most common RNA-binding protein families, as illustrated here, demonstrate the variability in the number of copies (as many as 14 in vigilin) and arrangements that exist. This variability has direct functional implications. For example, Dicer and RNase III both contain an endonuclease catalytic domain followed by a double-stranded RNA-binding domain (dsRBD). So, both proteins recognize dsRNA, but Dicer has evolved to interact specifically with RNA species that are produced through the RNA interference pathway through additional domains that recognize the unique structural features of these RNAs. Different domains are represented as coloured boxes. These include the RNA-recognition motif (RRM; by far the most common RNA-binding protein module), the K-homology (KH) domain (which can bind both single-stranded RNA and DNA), the dsRBD (a sequence-independent dsRNA-binding module) and RNA-binding zinc-finger (ZnF) domains. Enzymatic domains and less common functional modules are also shown. PABP, poly(A)-binding protein; PTB, polypyrimidine-tract binding; R/S, Arg/Ser-rich domain; SF1, splicing factor-1; TTP, tristetraprolin; U2AF, U2 auxiliary factor.

RNA recognition by RNA-binding modules

RNA-recognition motif. The RNA-recognition motif (RRM, also known as the RBD or the RNP motif) is by far the most common and best characterized of the RNA-binding modules. In this review, we will refer to it as RRM and use the term RBD for any domain that binds to RNA. The RRM is composed of 80–90 amino acids that form a four-stranded anti-parallel β-sheet with two helices packed against it, giving the domain the split αβ (βαββαβ) topology[18] (FIG. 3a). More than 10,000 RRMs have been identified that function in most, if not all, post-transcriptional gene-expression processes; in humans, ~0.5–1% of genes contain an RRM, often in multiple copies in the same polypeptide[19].

In the approximately 20 known structures of RRM–RNA complexes, RNA recognition usually occurs on the surface of the β-sheet[13–16,18,20–28]. Binding is mediated in most cases by three conserved residues: an Arg or Lys residue that forms a salt bridge to the phosphodiester backbone and two aromatic residues that make stacking interactions with the nucleobases. These three amino acids reside in the two highly conserved motifs, RNP motif-1 (RNP1) and RNP2, and define these motifs at the sequence level and are located in the two central β-strands[18]. This conserved platform allows for the recognition of two nucleotides in the centre of the β-sheet and of two additional nucleotides on either side[6]. However, a single RRM can recognize anywhere from four to eight nucleotides by using exposed loops and additional secondary structure elements that are not present in the canonical structure[3,6]. This general mechanism of recognition is found in many RRMs, but not in all[22,28]; some of these domains even interact with proteins and not with RNA[29–35]. So, some individual RRMs can bind to RNA with great specificity, but multiple domains are often needed to define specificity because the number of nucleotides that are recognized by an individual RRM is generally too small to define a unique binding sequence[3].

K-homology domain. The heterogeneous nuclear (hn)RNP K-homology domain (KH domain) is a domain that binds to both single-stranded (ss)DNA and ssRNA[36–42] and is ubiquitous in eukaryotes, eubacteria and archaea[43]. The domain is composed of ~70 amino acids with a functionally important signature sequence of (I/L/V)IGXXGXX(I/L/V) near the centre of the domain. Mutations in this region of the FMR1 protein cause fragile-X mental retardation syndrome[44]. All KH domains form a three-stranded β-sheet packed against three α-helices, but KH domains can be separated into two subfamilies on the basis of their topology[45] (type I has βααββα topology; type II has αββααβ topology). For both classes, four nucleotides are recognized in a cleft that is formed by the GXXG loop, the flanking helices, the β-strand that follows α2 (type I) or α3 (type II) and the variable loop between β2 and β3 (type I) or between α2 and β2 (type II; FIG. 3b). Unlike the RRM, this binding platform is free of aromatic amino acids; recognition is achieved instead by hydrogen bonding, electrostatic interactions and shape complementarity.

Double-stranded RBD. The double-stranded (ds)RBD is another small αβ domain of 70–90 amino acids that is found in both bacteria and eukaryotes. However, it interacts with dsRNA without making specific contacts with the nucleobases. The RNA-binding protein binds across two successive minor grooves and the intervening major groove on one face of the dsRNA helix[46] (FIG. 3c). Unlike the RRM or KH domains, the majority of the intermolecular contacts are sequence independent and involve 2′-OH groups and the phosphate backbone[46]. The presence of multiple dsRBDs can impart specificity for certain structures because of their ability to recognize certain arrangements of RNA helices[47–49]. In addition, the specificity of at least some dsRBDs is mediated

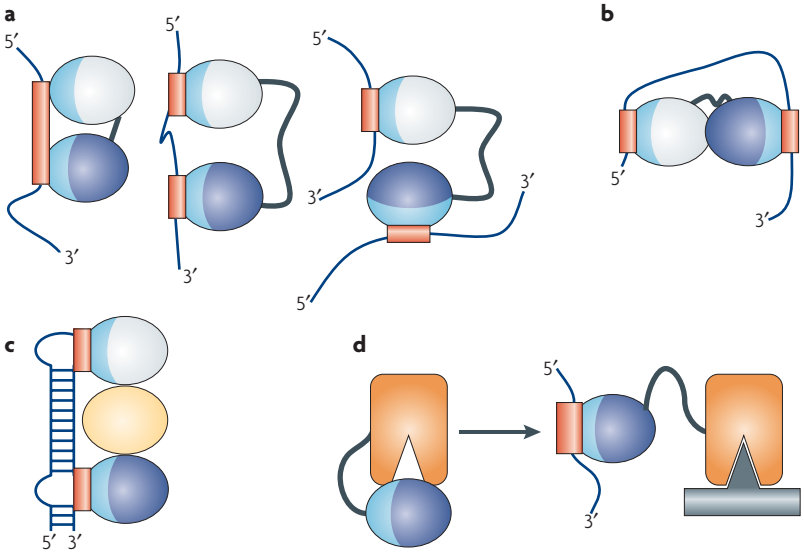

Figure 2 | **RNA-binding modules are combined to perform multiple functional roles.** RNA-binding domains (RBDs) function in various ways. **a** | They recognize RNA sequences with a specificity and affinity that would not be possible for a single domain or if multiple domains did not cooperate. Multiple domains combine to recognize a long RNA sequence (left), sequences separated by many nucleotides (centre), or RNAs that belong to different molecules altogether (right). **b** | RBDs can organize mRNAs topologically by interacting simultaneously with multiple RNA sequences. **c** | Alternatively, they can function as spacers to properly position other modules for recognition. **d** | They can combine with enzymatic domains to define the substrate specificity for catalysis or to regulate enzymatic activity. The RNA-binding modules are represented as ellipses with their RNA-binding surfaces coloured in light blue, and the corresponding binding sites in the RNA coloured in red; individual domains are coloured differently.

Zinc finger
A class of DNA- and RNA-binding proteins characterized by a Cys- and His-rich domain that chelates a zinc ion. Different classes of zinc-finger proteins contain different combinations of metal-binding amino acids: CCHH zinc fingers contain two Cys and two His residues, whereas CCCH and CCHC zinc-binding motifs contain three Cys and a single His residue in different topological arrangements.

AU-rich element
Sequences rich in A and U nucleotides that are found in the 3′ untranslated regions of mRNAs that promote stability or degradation of their associated RNAs, thus providing a mechanism for the control of gene expression.

Argonaute proteins
A family of proteins that are characterized by the presence of two homology domains, PAZ and PIWI. The proteins provide the essential catalytic activity for diverse RNA-silencing pathways.

in part by an N-terminal helix that binds to irregular helical elements in A-form RNA such as stem–loops, base mismatches and bulges[48,50–52] (FIG. 3c).

Zinc fingers. Zinc fingers are classical DNA-binding proteins that can also bind to RNA[53,54], as eloquently demonstrated by several recent structures[55–57]. They are typically classified on the basis of the residues that are used to coordinate zinc (Cys2His2 (CCHH), CCCH or CCHC) and are generally present in multiple repeats in a protein. Transcription factor TFIIIA (in which the motif was first identified) contains nine CCHH zinc fingers: fingers 1–3, 5 and 7–9 interact with DNA, whereas fingers 4–6 interact with 5S RNA[58,59] (finger 5 contacts both DNA and RNA). CCHH zinc fingers interact with DNA primarily by forming direct hydrogen bonds to Watson–Crick base pairs in the major groove, using residues in their recognition α-helix[60], whereas TFIIIA binds RNA by making specific contacts to two RNA loops through the recognition helices of fingers 4 and 6. So, zinc fingers can use some of the same residues to recognize both nucleic acids, but the different DNA and RNA structures dictate a distinct structural arrangement of the zinc fingers on the nucleic acid template.

A second family of RNA-binding zinc fingers contains CCCH motifs[61]. Remarkably, in the structure of the zinc-finger protein TIS11d bound to an AU-rich RNA element (ARE), sequence-specific RNA recognition occurs primarily through hydrogen bonding to the protein backbone[57] (FIG. 3d). So, the shape of the protein is the primary determinant of specificity as it provides a rigid hydrogen-bonding template. This mode of recognition is reminiscent of a third type of zinc fingers with a CCHC zinc-binding motif that is found in the nucleocapsid domain of the retroviral Gag proteins and in the HIV-1 nucleocapsid protein[62–63].

S1 domain. S1 domains were first identified in ribosomal protein S1 (hence the name), but have since been found in other RNA-binding proteins, including several exonucleases[64]. The domain is composed of ~70 amino acids arranged in a 5-stranded antiparallel β-barrel capped by a short 3_{10} helix, with 3 residues per turn[65]. The fold is similar to the oligonucleotide/oligosaccharide binding (OB)-fold superfamily, which also contains the related RNA-binding cold-shock domain[66]. The S1 domain uses the common OB-fold binding surface to recognize nucleic acids through two β-strands that are surrounded by several loops[67]. So, RNA binding by the S1 domain is reminiscent of RNA recognition by the RRM, in which a two-stranded β-sheet core contributes several conserved aromatic residues for stacking interactions with the nucleic acid bases, which are augmented by interactions provided by the surrounding loops and secondary structure elements[65,68].

PAZ and PIWI domains. RNA processing during RNA interference (RNAi) and micro (mi)RNA biogenesis generate species with unique structural and chemical features that must be recognized specifically, but in a sequence-independent manner. These functional requirements are fulfilled by a specialized set of domains that are encountered in proteins involved in processing miRNA and small interfering (si)RNA precursors.

The 110-amino-acid PAZ domain contains a β-barrel domain that resembles an OB or S1 fold juxtaposed to a small αβ domain that forms a clamp-like structure in which RNA binds[69–71] (TABLE 1). The PAZ domain selectively binds to the two-nucleotide overhangs and probably serves as an anchor to position the miRNA for proper cleavage by the RNase III-type nuclease Dicer[72,73]. PAZ domains in Argonaute proteins facilitate cleavage of the target strand by the RNA-induced silencing complex (RISC), which is responsible for degradation of RNA that is targeted for silencing. The additional PIWI domain in Argonaute instead adopts an RNase H fold and anchors the unique 5′ end of the guide strand to position the target strand for degradation[74–78] (TABLE 1).

Expanding conventional RNA-binding surfaces. The type of RNA that can be recognized by RBDs is increased not only by proteins with multiple domains (as discussed below), but also by expanding a canonical RNA-binding surface through additional secondary structures or loops[6,52]. In the reverse situation, a canonical recognition surface can be occluded by secondary structure elements, leading to the regulation of the RNA-binding activity. So, many proteins that are involved in spliceosome assembly have RNA-binding modules that differ from

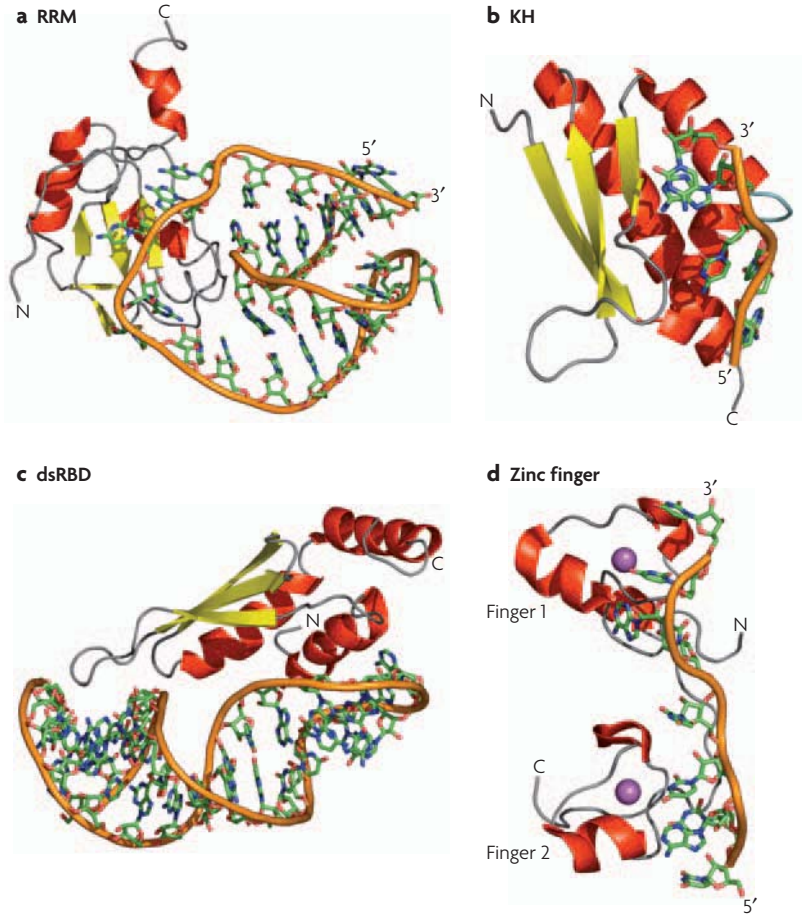

Figure 3 | **How RNA-binding modules recognize RNA. a** | Structure of the N-terminal RNA-recognition motif (RRM) of human U1A bound to RNA[18]. In this structure, and in many other RRM–RNA complexes, single-stranded bases are specifically recognized through the protein β-sheet and through two loops that connect the secondary structure elements. **b** | The K-homology-3 (KH3) domain of Nova-2 bound to 5′-AUCAC-3′[41]. KH domains bind to both single-stranded DNA and RNA through a conserved GXXG sequence that is located in an exposed loop (light blue). **c** | The yeast Rnt1 double-stranded RNA-binding domain (dsRBD) bound to an RNA helix capped by an AGNN tetraloop[52]. A conserved protein loop (left-most part of the structure) interacts with 2′-OH groups in the RNA minor groove, whereas highly conserved Lys and Arg residues at the end of the longer helix recognize the position of phosphate atoms that are characteristic of an A-form helix. **d** | The two zinc fingers of TIS11d bound to an AU-rich RNA element[57]. The identity of the single-stranded RNA is recognized by the protein backbone through hydrogen bonds with the Watson–Crick face of each base. In all panels, the RNA backbone is represented with an orange ribbon, α-helices are in red and β-sheets are in yellow; the zinc atom in the TIS11d structure is in magenta.

their canonical structure. For example, splicing factor-1 (SF1), which binds to the branch-point sequence, has an additional Quaking homology-2 (QUA2) domain that defines an enlarged KH domain by making extensive hydrophobic interactions with the KH domain itself. By increasing the recognition surface, SF1 can bind to the seven single-stranded nucleotides that define the branch-point sequence[42] (FIG. 4b).

The structures of the first two quasi-RRMs (qRRMs) from hnRNP F demonstrate instead how an RRM can use a different surface for RNA recognition when the β-sheet surface is occluded[79]. This member of the hnRNP family is involved in the recognition of G-rich sequences (G-tracts) that are often found at recognition elements that are responsible for 5′-splice-site recognition[80–82]. In the structure of the hnRNP F protein bound to the G-tract in *Bcl-x* pre-mRNA, each domain resembles a canonical RRM despite the absence of the RNP1 and RNP2 motifs normally used to bind RNA. Furthermore, the β-sheet surface is occluded by the presence of a C-terminal α-helix packed against it. So, the first two qRRMs of hnRNP F recognize RNA through a novel surface that is composed of a small β-hairpin between α2 and β4 and the β1–α1 and β2–β3 loops[79]. Perhaps the requirement for binding through a different surface in this complex stems from the necessity to recognize G-quadruplex RNA and at the same time to prevent nonspecific binding to ssRNAs, which are normally recognized by RRM proteins.

An additional α-helix that is C-terminal to the canonical domain is common in RRMs. The C-terminal domain of La protein, human cleavage-stimulation factor-64 (CSTF64) and U1A all have a helix at the C terminus of the domain[12,20,83] (FIG. 3a). Many other domains form such a helix when bound to RNA; for example, in yeast Hrp1, HuD and PABP[14,16,25]. The C-terminal RRM of La does not interact with RNA at all and, in the U1A and CstF64 structures, the helix moves away from the β-sheet to allow RNA recognition by the canonical site (FIG. 3a), which suggests that these helices primarily perform a regulatory role.

Multiple domains specify RNA recognition

Tandem domains. Isolated RBDs generally have limited ability to interact with RNA in a sequence-specific manner because their recognition sequences are too short[6]. Multiple domains (most typically two) are therefore tethered together on a single polypeptide to create a much larger binding interface that recognizes a longer sequence. Perhaps the most extreme example of this concept comes from the Pumilio (Puf) family of proteins. Each domain recognizes a single nucleotide on its own, but by combining multiple repeats the protein can bind with high affinity and specificity to as many as eight nucleotides[84] (TABLE 1; FIG. 5a). In fact, the three amino acids that recognize a particular nucleotide provide a reasonably predictive recognition code that can be exploited to engineer proteins that recognize different RNA sequences from those specified by the wild-type proteins[84,85].

Interdomain arrangement. Multiple domains associate with each other in various ways to generate extended RNA-recognition interfaces. The recent structure of Hrp1 (FIG. 5b) exemplifies the structural principles that are involved in RNA recognition by two RRMs in tandem, which are also observed in Sex-lethal (Sxl), PABP, nucleolin and HuD proteins[13–15,25]. In the free protein, both domains function as independent, rigid structures separated by a short flexible linker. Upon binding, both protein and RNA undergo significant changes in structure, with the linker forming a short helix and several interdomain contacts, which creates a compact surface for the recognition of adjacent stretches in the RNA[16] (FIG. 5b).

RNA-induced silencing complex
A multicomponent ribonucleoprotein complex that cleaves specific mRNAs that are targeted for degradation by homologous double-stranded RNAs during the process of RNA interference.

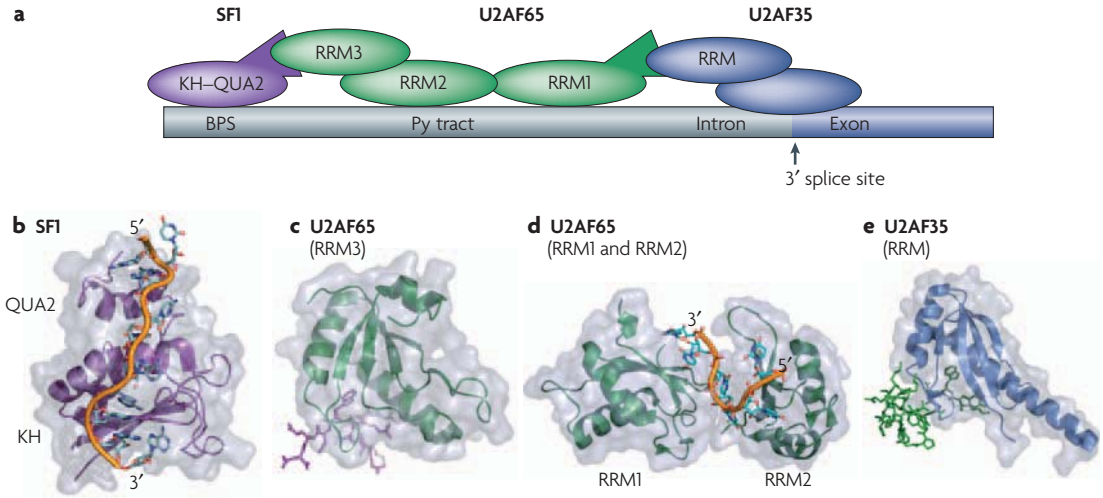

Figure 4 | **Protein–protein interactions and protein-RNA interactions define the site of spliceosomal assembly.** **a** | Schematic of the interactions between various proteins and RNA at the splicing site. The structures of some of the key domains that are involved in these interactions are shown in panel **b**. In the RNA, the branch-point sequence (BPS), pyrimidine tract (Py tract), and the 3′ splice site are labelled with the intron shown in grey and the exon in dark blue. **b** | Splicing factor-1 (SF1) recognizes the BPS through its K-homology (KH)–Quaking homology-2 (QUA2) domains, which creates an extended KH domain that can recognize the full BPS sequence RNA[42]. **c** | This interaction is strengthened by protein–protein interactions between the N terminus of SF1 and the non-canonical RNA-recognition motif-3 (RRM3) of U2 auxiliary factor subunit-65 (U2AF65)[35]. **d** | RRM3 is bound to the pyrimidine tract through its first two canonical RRMs, RRM1 and RRM2 (REF. 10). **e** | Last, the U2AF65 interaction is also aided by protein–protein interactions between its N terminus and the non-canonical RRM of U2AF35 bound at the 3′ splice site[33]. The protein and peptide structures are colour coded as in panel **a**.

By contrast, when TIS11d binds to AREs, there are few interdomain interactions. However, a pre-organized linker between the two zinc fingers orients the two domains for recognition of an eight-nucleotide RNA by the protein main chain with little side-chain involvement[57] (FIG. 3d). In a third example, the two KH domains in the structure of the transcription factor NusA bound to RNA make extensive interdomain contacts with each other[86]. This association of the KH domains creates an extended RNA-binding surface that allows the recognition of an 11-nucleotide RNA[37] (FIG. 5c). Each of the KH domains of NusA specifically recognizes four nucleotides, as is expected for KH domains; their separation by a three-nucleotide linker that also makes interactions with the protein generates the complete recognition sequence[37]. In this structure, the binding interface is further extended by an S1 domain that is N-terminal to the first KH domain and that makes extensive interdomain contacts and, in doing so, can provide an additional surface for RNA recognition.

The zinc-finger domains of TFIIIA provide yet another example of how linkers between RNA-recognition domains play a crucial part in substrate recognition: the linker in this case is a zinc-finger module. In the TFIIIA–5S RNA complex, fingers 4 and 6 interact extensively with the RNA, whereas finger 5 functions as a spacer that makes sequence-independent contacts that involve the side chains of its α-helix and the RNA backbone. Effectively, finger 5 serves as a bridge between loops E and A in 5S RNA, which are directly recognized by fingers 4 and 6, respectively[56] (FIG. 5f).

Pumilio (Puf) family of proteins
A highly conserved family of RNA-binding proteins with a C-terminal RNA-binding domain that is composed of eight tandem repeats, with each repeat recognizing a single nucleotide.

Although the previous examples illustrate the importance of an ordered linker, the presence of a long flexible linker can be favoured at times (FIG. 2a) because it allows RNA-binding proteins to recognize sites that have a variable number of nucleotides between them, or that are separated from each other on the same RNA or on different RNA molecules altogether. In all of these cases, ordering of the linker upon binding RNA is not likely to occur. A good illustration of this situation is provided by the two dsRBDs of the RNA-editing enzyme ADAR2, in which the two domains do not interact and are separated by a flexible linker in the free or bound protein[48] (FIG. 5d). As ADAR2 is required to edit multiple RNAs, interdomain flexibility allows each dsRBD to bind to its preferred site in RNAs of varying length and structure.

Yet another example of the potential advantages of connecting domains with flexible linkers can be found in complexes in which conformational flexibility is required for function. In the complex between the far-upstream element (FUSE) and the FUSE-binding protein (FBP), a 30-residue linker separates the KH3 and KH4 domains of FBP so that they can move independently of each other even when the protein is bound to DNA[39]. This property is likely to be functionally important because FBP binds to and modulates the helicase activity of the general transcription factor TFIIH. As this protein might function as a torque-generating machine, it is important for FBP to bind to the dynamic TFIIH molecule while maintaining its interaction with DNA.

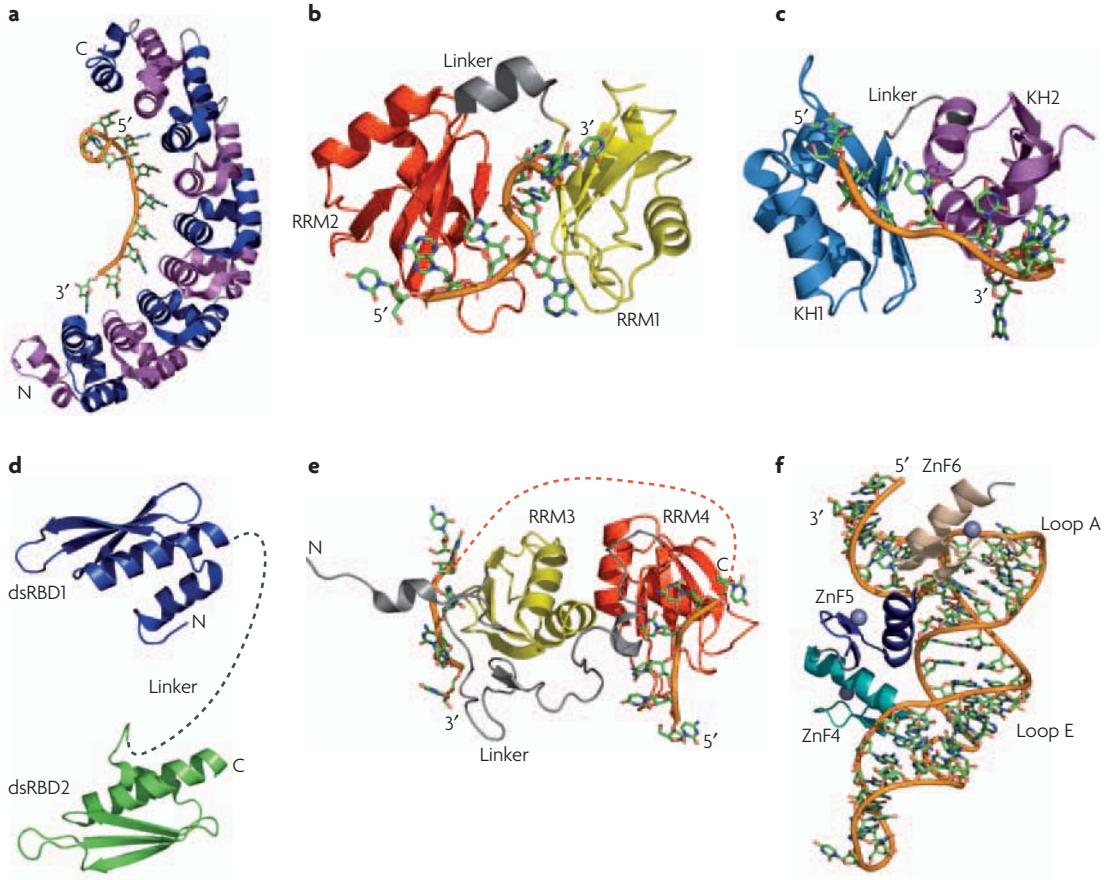

Figure 5 | **RNA-binding modules function together to recognize a specific RNA. a** | The structure of human Pumilio protein provides an example of how multiple repeats (eight in this case) that individually recognize a few nucleotides (one in this case) combine to specifically recognize a much longer RNA sequence. Repeats are alternatively coloured in magenta and blue; the RNA is coloured similarly in all other structures with the backbone shown in orange. **b**,**c** | In the structures of the two RNA-recognition motifs (RRMs) from yeast Hrp1 (REF. 16) (panel **b**) and the two K-homology (KH) domains of NusA[37] (panel **c**), a short linker (grey) allows the two domains to position themselves with respect to one another upon binding RNA. For Hrp1, RRM1 is yellow and RRM2 is red; for NusA, KH1 is cyan and KH2 is purple. **d** | Flexibility in the linker between two double-stranded RNA-binding domains (dsRBDs) allows the recognition of separated binding sites. The two dsRBDs of ADAR2 are connected by a flexible linker (dashed line) that can allow the protein to interact with various targets of different structure[48]. **e** | RRM3 (yellow) and RRM4 (red) of polypyrimidine-tract binding (PTB) protein form interdomain interactions that involve the face of the protein opposite to the β-sheet that is important for RNA recognition. This interaction positions the two domains in such a way that interacting RNA sequences are looped away from each other, as indicated by the orange dotted line that connects the two RNAs[28]. **f** | The structure of the TFIIIA–RNA complex illustrates how zinc finger 5 (ZnF5; dark blue) functions as a spacer that positions zinc fingers 4 (teal) and 6 (tan) for recognition of loops E and A, respectively, in 5S ribosomal RNA[55].

This theme is observed even in proteins that contain RRM domains, a departure from the common and canonical arrangement described above for yeast Hrp1 and other proteins[13–16,25]. The structure of polypyrimidine-tract binding (PTB) protein shows that the RRM3 and RRM4 are connected by a long linker and interact with each other in a way that forces their respective RNA-binding surfaces to face in opposite directions[28]. This orientation is essentially the opposite of what is observed in many di-RRM proteins, yet might be functionally crucial in splicing regulation by causing the exon or branch-point sequences to loop out, preventing binding of spliceosomal components and repressing splicing (FIG. 5e).

The linker length is important. The considerations mentioned above indicate that one of the major determinants for the affinity and specificity of RNA-binding proteins that contain multiple domains resides in the amino acids that link the domains. The length and rigidity of the linker can have dramatic effects on RNA affinity[87] and can influence whether a protein binds a single RNA or multiple RNAs (FIG. 2a, right). Using the assumption that the free energy of binding individual domains is additive, we would expect the affinity of a protein with multiple RBDs to be the product of the affinity of the individual domains. However, when the linker remains flexible (for example, in hnRNP A1) the affinity of the two-domain protein is much less (1,000-fold for hnRNP A1) than

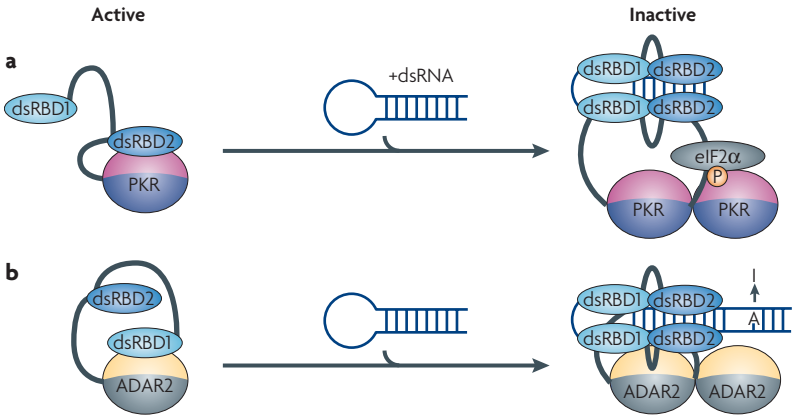

Figure 6 | **Modular architecture allows for the regulation of the catalytic activity of enzymatic domains.** In both the protein kinase PKR and the ADAR RNA-editing proteins, interdomain interactions between an RNA-binding module and a catalytic domain maintain the proteins in an inactive state. **a** | The kinase domain of PKR is inhibited by an interaction with the double-stranded RNA-binding domain-2 (dsRBD2). Binding to dsRNA releases the kinase from its inactive state, allowing it to inhibit translation by phosphorylating the α-subunit of eukaryotic initiation factor-2α (eIF2α). **b** | The activity of ADAR2 is controlled by a mechanism similar to that which controls PKR, but in this case dsRBD1 is involved in the inactivation of the catalytic domain. When dsRNA binds to both dsRBDs, the protein dimerizes and the catalytic domain becomes exposed to convert adenosine to inosine. P, phosphate.

the product of the affinities of the individual domains[88]. When the first RBD is bound, the second domain sweeps a volume that is proportional to the length of the linker. In this sphere, the effective concentration of the second domain is different than in the free solution, leading to altered affinity. A simple model was developed to calculate how the length of the linker affects affinity; using this model, long linkers (>50–60 residues) are predicted to have a negligible effect on affinity because the two domains function independently of each other. As the linker gets shorter, the affinity for RNA increases between 10- and 1,000-fold when compared with the affinity of individual RRMs added together[87].

This simple model assumes that the linker does not contact the RNA, but in many cases the linker becomes ordered after binding RNA. In the example of the nucleolar phosphoprotein nucleolin, the model would predict a 100-fold increase in affinity compared with that of the two individual RRMs of nucleolin, but an increase of between 1,000- and 100,000-fold was observed, depending on the RNA sequence that was tested[89]. Part of the increase in affinity was attributable to the ordering of the linker into an α-helix that effectively shortens its length by half. When the prediction was repeated with this correction, predicted and measured affinities agreed to within 10-fold for some RNAs. However, because of direct interactions between the linker and target RNAs, even this calculation could not account for the 1,000-fold difference between the predicted and observed affinities for other RNAs[89].

Protein–protein interactions

Dimerization of RNA-binding proteins. In addition to expanding the ways in which RNA can be recognized, multiple modules also allow RNA-binding proteins to interact simultaneously with other proteins and with RNA. The simplest example of this is dimerization. Two proteins that are involved in the viral response to RNA silencing provide exquisite examples of how dimerization allows specific interactions to be established that would not be possible in the isolated proteins.

The p19 protein is required for tombusvirus virulence in plants, and can also suppress the RNAi response when expressed in *Drosophila melanogaster* and human cells[90,91]. It functions by specifically binding to siRNAs and preventing their loading into the RISC complex[92]. Two structures of p19 proteins bound to 21-nucleotide siRNA demonstrate that the protein adopts an αβ topology and binds RNA as a homodimer. The RNA-binding surface is formed by a continuous eight-stranded β-sheet that is formed by the two monomers that flank each end by an α-helix. Each monomer measures the length of the siRNA by providing Trp residues in this α-helix that form stacking interactions with the bases at the 5′ and 3′ end of the siRNA; so, the position of the Trp is defined by the structure of the homodimer. Dimerization of p19 allows this protein to measure the length of the siRNA with great precision by positioning the two critical Trp side chains[92,93].

Another potent viral suppressor of RNAi is the Flock House virus B2 protein. Its structure is composed of three α-helices that dimerize to create a four-helix bundle that recognizes RNA along one face of an A-form helix[94,95]. Structural and biochemical evidence demonstrated that this protein suppresses silencing in two ways: by binding to siRNAs and thereby preventing loading into RISC, and by coating longer dsRNA precursors and protecting them from cleavage by Dicer. For both p19 and Flock House virus B2, the conserved features of the siRNAs (their size and double-helical character)[92–95] are recognized because dimerization generates extended binding sites out of small protein domains and because it establishes the relative position of amino acids that are involved in RNA recognition.

These two examples illustrate the role of dimerization in RNA recognition, but there are other examples of RBDs that function by dimerizing or by forming protein–protein interactions. In the structure of the N-terminal RRM of U1A bound to an RNA regulatory element in its own 3′ untranslated region (3′ UTR), two separate RRMs interact through their C-terminal helices to form a homodimer after binding to the RNA. This cooperative binding event can only occur in the presence of RNA because the C-terminal helix is associated with the β-sheet surface of the RRM in the free protein. Dimerization also creates an interface that inhibits polyadenylation by direct interaction with poly(A) polymerase[24]. In the KH3 domain of Nova-1, changes in the rigidity of the protein are observed upon dimerization, and this stiffening of the entire protein can help in nucleic acid recognition by reducing the entropic cost of binding to RNA. Furthermore, dimerization presents two recognition sites for RNA binding and can therefore provide a cooperative interaction that strengthens the affinity of the protein for the RNA[96].

The formation of heterodimers through interactions between an RBD and another protein can increase the specificity of the RNA interaction as well. For example, the binding of the RRM of spliceosomal U2B′′ to a stem–loop in U2 small nuclear (sn)RNA requires an interaction with the U2A′ protein[23]. In a different example, the cap-binding protein-80 (CBP80) subunit of the cap-binding complex must interact with the RRM of the CBP20 subunit if this RRM is to bind with high affinity to the 7-methylguanosine cap of mRNA[22,97]. The recent structures of the archaeal and eukaryotic exosomes have revealed extensive protein–protein interactions between proteins that contain both KH and S1 domains in the core of the protein complex[98,99]. These interactions can position the S1 domains of specific exosome subunits to recognize the RNAs that are targeted for degradation.

Protein–protein interactions define RNA specificity. RBDs from different proteins can cooperate to recognize RNA through a combination of weak protein–RNA and protein–protein interactions. The recent dissection of a complex that was derived from the spliceosome demonstrates this principle and illustrates how even small sequence and structural alterations in RBDs can modulate the RNA-recognition properties of RBDs indirectly by altering protein–protein interactions.

During initial steps in spliceosome assembly, SF1 and U2AF proteins cooperatively bind to sequences at the 3′ splice site and upstream of it[100] (FIG. 4a). Recognition of RNA *cis*-acting elements by the two U2AF subunits, U2AF65 and U2AF35, commits the pre-mRNA to the splicing reaction[101]. Specifically, U2AF65 recognizes the polypyrimidine tract in the pre-mRNA primarily through its two central canonical RRMs[10,102] (FIG. 4a,d); this interaction is strengthened by the interaction between a third non-canonical RRM in this protein and SF1 (FIG. 4a,c)[35], which is bound at the branch-point sequence through a KH domain (FIG. 4a,b). Additional cooperativity in the assembly of this complex is provided by protein–protein interactions between a non-canonical RRM in U2AF35 (FIG. 4a,e), bound at the 3′ splice site, and the N terminus of U2AF65 (REF. 33).

Protein–protein interaction surfaces. As described above, RRM domains can form protein–protein as well as protein–RNA interactions. The protein–protein interactions occur through non-canonical RRM domains in both U2AF65 and U2AF35, which have a much longer α1 helix compared with other RRMs, and these helices are the primary mediators of the protein–proteins interactions observed in this complex[33,35] (FIG. 4c,e). Closer inspection of the U2AF structures reveal common themes that might indicate whether an RRM binds to a protein: these themes include poor conservation of the RNP motifs, an Arg-X-Phe motif in the last loop of the RRM and conserved acidic residues in the α1 helix[103]. These features define a novel functional class, the U2AF-homology motifs (UHMs), that can form protein–protein interactions.

The UHM class does not exhaust all possible ways in which two RRMs can interact. The interactions of other RRMs with proteins (for example, the human Y14–Magoh structure from the exon-junction complex and the human UPF2–UPF3 RNA-surveillance complexes[29,31,32,34,104,105]) occur on the surface of the β-sheet through residues that are involved in RNA binding in other RRMs. Until more structures of such protein–protein complexes become available, the sequence and structural features in such RRMs that allow them to bind to other proteins, rather than to RNA, will remain unclear.

RBDs other than the RRM can participate in protein–protein interactions as well. As previously described, a number of KH domains can dimerize, and dsRBD domains form protein–protein interactions that regulate the assembly of complexes that are involved in RNA localization as well as the catalytic activity of enzymes that function on dsRNA. For example, Staufen, a protein that is involved in RNA localization in early development and in neurons, contains up to five dsRBDs. Some domains can bind dsRNA[51], whereas other domains bind other proteins during embryogenesis[106]. Remarkably, surface-exposed amino acids that are involved in RNA recognition are conserved among dsRBDs of Staufen that bind to dsRNA, but not in protein-binding dsRBDs. For these domains, it is the surface opposite to the dsRNA in the canonical dsRBD–dsRNA structure that is conserved instead[51]. So, the ability of a protein to bind to other proteins can be as important functionally as its RNA-binding activity.

Catalytic domains that function on RNA

Positioning catalytic domains onto their substrate. Modularity allows RBDs to target a substrate and to promote or repress the enzymatic activity of catalytic domains in the same polypeptide (FIG. 2d). The way in which RNA-binding and enzymatic modules are positioned in a protein can define how a particular protein recognizes RNA. Furthermore, the enzymatic activity can also be enhanced or repressed through mutually exclusive or cooperative interactions between RBDs, catalytic domains and RNA.

An elegant example of how domain positioning facilitates enzymatic function comes from the RNAi pathway. In the first step of the cascade that leads to gene silencing, the nuclear enzymes Drosha and Pasha process primary miRNAs to stem–loops of ~70 nucleotides; Dicer subsequently binds to these miRNA precursors by recognizing two 3′-terminal nucleotide overhangs that are generated by Drosha[107]. A minimal Dicer structure from *Giardia intestinalis* (lacking the N-terminal helicase and the C-terminal dsRBD; FIG. 1) demonstrates that Dicer probably functions as a molecular ruler that positions the catalytic RNase III domains ~25 nucleotides from where the 3′ overhanging nucleotides are recognized by its PAZ domain[72], which corresponds to the approximate length of siRNAs.

Another particularly beautiful example of this principle is found in the recent structure of a complete archaeal box H/ACA small nucleolar RNP (snoRNP)[108]. These particles are responsible for the catalytic conversion of uracil to pseudouridine in ribosomal and other RNAs[109]. In this structure, the site of pseudo-uridylation is juxtaposed to the catalytic centre of the protein enzyme

Exosome
A multisubunit 3′→5′ exonuclease that functions in the nucleus and the cytoplasm in several different RNA-processing and RNA-degradation pathways.

Exon-junction complex
A multisubunit protein complex that is deposited on the mRNA during the splicing reaction near the splice site. It remains bound to the RNA during subsequent gene-expression events, and serves as a platform to recruit nuclear and cytoplasmic factors that influence mRNA localization, transport, stability and translation.

Cbf5 (also known as dyskerin) by two protein clamps at either end of the RNA. The 3′-terminal 'clamp' (the ACA sequence motif that defines this class of non-coding RNAs) is recognized by the PUA domain of Cbf5, whereas the second clamp (the apical loop of the non-coding RNA) is recognized by a complex of Cbf5 with two other protein components of the particle.

Enzyme activation and repression in response to RNA. The dsRNA-dependent protein kinase PKR (FIG. 6a) and the RNA-editing enzyme ADAR2 (FIG. 6b) provide examples of how RBDs can modulate enzymatic activity by interacting with both the substrate RNA and the catalytic domain (FIG. 2d). PKR is an interferon-induced kinase that has a key role in controlling viral infection and in maintaining cellular homeostasis by becoming activated in response to double-stranded viral RNAs. In the active form, it phosphorylates the α-subunit of eukaryotic initiation factor-2 (eIF2), thereby inhibiting translation and suppressing viral spread[110]. ADARs function instead on dsRNA to catalyse the conversion of adenosine to inosine, which is then recognized as guanosine, affecting both the primary sequence and structure of the edited RNA[111].

Both proteins have two N-terminal dsRBDs that bind to dsRNA; in each case, the dsRBDs function both as RNA-recognition units and as auto-inhibitors of the catalytic domain[112,113]. In PKR, the second dsRBD masks the kinase domain by binding to it directly, thereby maintaining its inactive state[112,114,115] (FIG. 6a). In ADAR2, the proposed inhibitory element is the first dsRBD[113] (FIG. 6b). In both proteins, RNA binding causes enzyme activation by relieving the auto-inhibition caused by the interactions between the RNA-binding and catalytic domains. As both ADAR and PKR require RNA of sufficient length for activation, the two dsRBDs might be necessary to fully de-repress the catalytic activity[113]. In PKR, the presence of a sufficiently long dsRNA (for example, viral RNAs such as HIV TAR) allows both dsRBDs to cooperatively bind to RNA[116,117], relieving the structural block and allowing the kinase domain to be activated through autophosphorylation and dimerization[118–120]. The initial event in this cascade is likely to be the binding of the first dsRBD to dsRNA, because this domain has much higher affinity for RNA than does the second domain[117]. Only in the presence of a sufficiently long dsRNA can the second dsRBD bind as well, thereby releasing the kinase from its inactive state.

Conclusions

Many RNA-binding proteins are composed of a few modules of conserved structure but of often limited sequence specificity. By combining these motifs in various structural arrangements, evolution has generated proteins that can recognize RNA with the affinity and selectivity that is required to find cognate RNAs in the cellular medium, while retaining the versatility required to regulate, assemble and disassemble RNA-processing complexes. Structural biology has provided the molecular details about how individual domains recognize RNA, but many of these proteins require multiple copies of one of several common domains to function (FIG. 1). It is therefore important to understand how multiple modules bind RNA, and how the modular nature of these proteins specifies their biological function. We have described some of the structural principles of how multiple domains recognize an RNA (FIG. 2), but there are still only a few structures of proteins that contain multiple RBDs. Recent studies have also led to the observation that RNA-binding modules can regulate the biological activity of enzymes that act on RNAs in ways that go beyond the identification of the target RNA, but full understanding of these regulatory mechanisms will require detailed structural characterization that is not yet available. We expect that future structural analyses will expand on the diverse ways in which combinations of RBDs can augment protein function.

1. Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nature Rev. Mol. Cell Biol.* **3**, 195–205 (2002).
2. Burd, C. G. & Dreyfuss, G. Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265**, 615–621 (1994).
3. Auweter, S. D., Oberstrass, F. C. & Allain, F. H. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.* **34**, 4943–4959 (2006).
This review provides a comprehensive analysis of several RBDs and uses the recognition principles discovered in the past 10 years to construct a set of rules for RNA recognition by each domain.
4. Chang, K. Y. & Ramos, A. The double-stranded RNA-binding motif, a versatile macromolecular docking platform. *FEBS J.* **272**, 2109–2117 (2005).
5. Hall, T. M. Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.* **15**, 367–373 (2005).
6. Maris, C., Dominguez, C. & Allain, F. H. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **272**, 2118–2131 (2005).
7. Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445–452 (2003).
8. Doolittle, R. F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314 (1995).
9. Bork, P., Downing, A. K., Kieffer, B. & Campbell, I. D. Structure and distribution of modules in extracellular proteins. *Q. Rev. Biophys.* **29**, 119–167 (1996).
10. Sickmier, E. A. *et al.* Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Mol. Cell* **23**, 49–59 (2006).
11. Deka, P., Rajan, P. K., Perez-Canadillas, J. M. & Varani, G. Protein and RNA dynamics play key roles in determining the specific recognition of GU-rich polyadenylation regulatory elements by human Cstf-64 protein. *J. Mol. Biol.* **347**, 719–733 (2005).
12. Perez Canadillas, J. M. & Varani, G. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J.* **22**, 2821–2830 (2003).
13. Allain, F. H., Bouvet, P., Dieckmann, T. & Feigon, J. Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. *EMBO J.* **19**, 6870–6881 (2000).
14. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98**, 835–845 (1999).
15. Handa, N. *et al.* Structural basis for recognition of the tra mRNA precursor by the sex-lethal protein. *Nature* **398**, 579–585 (1999).
This was among the first reports to describe what is now a common mode of RNA recognition: proteins that contain tandem RRM domains.
16. Perez-Canadillas, J. M. Grabbing the message: structural basis of mRNA 3′ UTR recognition by Hrp1. *EMBO J.* **25**, 3167–3178 (2006).
17. Birney, E., Kumar, S. & Krainer, A. R. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res.* **21**, 5803–5816 (1993).
18. 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* **372**, 432–438 (1994).
19. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).
20. Allain, F. H. *et al.* Specificity of ribonucleoprotein interaction determined by RNA folding during complex formulation. *Nature* **380**, 646–650 (1996).
21. Ding, J. *et al.* Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes Dev.* **13**, 1102–1115 (1999).
22. Mazza, C., Segref, A., Mattaj, I. W. & Cusack, S. Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. *EMBO J.* **21**, 5548–5557 (2002).
23. 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* **394**, 645–650 (1998).

24. Varani, L. *et al.* 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. *Nature Struct. Biol.* **7**, 329–335 (2000).
25. Wang, X. & Tanaka Hall, T. M. Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nature Struct. Biol.* **8**, 141–145 (2001). **Reports the first evidence that RRMs bind RNA using a recognition mode that is highly conserved in proteins that contain single or multiple domains, providing a structural code for recognition.**
26. Auweter, S. D. *et al.* Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* **25**, 163–173 (2006).
27. Hargous, Y. *et al.* Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. *EMBO J.* **25**, 5126–5137 (2006).
28. Oberstrass, F. C. *et al.* Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309**, 2054–2057 (2005). **The structure of all four RRMs from PTB bound to RNA provide insight into the diverse ways in which even related domains form different RNA-recognition platforms by interacting with other RRMs in different ways.**
29. Bono, F., Ebert, J., Lorentzen, E. & Conti, E. The crystal structure of the exon junction complex reveals how it maintains a stable grip on mRNA. *Cell* **126**, 713–725 (2006).
30. Bono, F. *et al.* Molecular insights into the interaction of PYM with the Mago–Y14 core of the exon junction complex. *EMBO Rep.* **5**, 304–310 (2004).
31. Fribourg, S., Gatfield, D., Izaurralde, E. & Conti, E. A novel mode of RBD-protein recognition in the Y14–Mago complex. *Nature Struct. Biol.* **10**, 433–439 (2003).
32. Kadlec, J., Izaurralde, E. & Cusack, S. The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. *Nature Struct. Mol. Biol.* **11**, 330–337 (2004).
33. 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 U2AF35/U2AF65 heterodimer. *Cell* **106**, 595–605 (2001).
34. 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.* **13**, 933–941 (2003).
35. Selenko, P. *et al.* Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol. Cell* **11**, 965–976 (2003).
36. Backe, P. H., Messias, A. C., Ravelli, R. B., Sattler, M. & Cusack, S. X-ray crystallographic and NMR studies of the third KH domain of hnRNP K in complex with single-stranded nucleic acids. *Structure* **13**, 1055–1067 (2005).
37. Beuth, B., Pennell, S., Arnvig, K. B., Martin, S. R. & Taylor, I. A. Structure of a *Mycobacterium tuberculosis* NusA–RNA complex. *EMBO J.* **24**, 3576–3587 (2005).
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.* **21**, 3476–3485 (2002).
39. Braddock, D. T., Louis, J. M., Baber, J. L., Levens, D. & Clore, G. M. Structure and dynamics of KH domains from FBP bound to single-stranded DNA. *Nature* **415**, 1051–1056 (2002).
40. Du, Z. *et al.* Crystal structure of the first KH domain of human poly(C)-binding protein-2 in complex with a C-rich strand of human telomeric DNA at 1.7 Å. *J. Biol. Chem.* **280**, 38823–38830 (2005).
41. Lewis, H. A. *et al.* Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* **100**, 323–332 (2000).
42. Liu, Z. *et al.* Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294**, 1098–1102 (2001).
43. Siomi, H., Matunis, M. J., Michael, W. M. & Dreyfuss, G. The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. *Nucleic Acids Res.* **21**, 1193–1198 (1993).
44. De Boulle, K. *et al.* A point mutation in the *FMR-1* gene associated with fragile X mental retardation. *Nature Genet.* **3**, 31–35 (1993).
45. Grishin, N. V. KH domain: one motif, two folds. *Nucleic Acids Res.* **29**, 638–643 (2001).
46. Ryter, J. M. & Schultz, S. C. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* **17**, 7505–7513 (1998).
47. Stephens, O. M., Haudenschild, B. L. & Beal, P. A. The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity. *Chem. Biol.* **11**, 1239–1250 (2004).
48. Stefl, R., Xu, M., Skrisovska, L., Emeson, R. B. & Allain, F. H. Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. *Structure* **14**, 345–355 (2006).
49. Xu, M., Wells, K. S. & Emeson, R. B. Substrate-dependent contribution of double-stranded RNA-binding motifs to ADAR2 function. *Mol. Biol. Cell* **17**, 3211–3220 (2006).
50. Leulliot, N. *et al.* A new α-helical extension promotes RNA binding by the dsRBD of Rnt1p RNAse III. *EMBO J.* **23**, 2468–2477 (2004).
51. Ramos, A. *et al.* RNA recognition by a Staufen double-stranded RNA-binding domain. *EMBO J.* **19**, 997–1009 (2000).
52. Wu, H., Henras, A., Chanfreau, G. & Feigon, J. Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Proc. Natl Acad. Sci. USA* **101**, 8307–8312 (2004).
53. Carballo, E., Lai, W. S. & Blackshear, P. J. Feedback inhibition of macrophage tumor necrosis factor-α production by tristetraprolin. *Science* **281**, 1001–1005 (1998).
54. Picard, B. & Wegnez, M. Isolation of a 7S particle from *Xenopus laevis* oocytes: a 5S RNA–protein complex. *Proc. Natl Acad. Sci. USA* **76**, 241–245 (1979).
55. Lee, B. M. *et al.* Induced fit and "lock and key" recognition of 5S RNA by zinc fingers of transcription factor IIIA. *J. Mol. Biol.* **357**, 275–291 (2006).
56. Lu, D., Searles, M. A. & Klug, A. Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature* **426**, 96–100 (2003). **This structure provides the first example of a zinc-finger protein bound to RNA, and also shows how an entire domain can function as a linker to position zinc fingers 4 and 6 for recognition of their respective binding sites and space them as needed.**
57. Hudson, B. P., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nature Struct. Mol. Biol.* **11**, 257–264 (2004).
58. Clemens, K. R. *et al.* Molecular basis for specific recognition of both RNA and DNA by a zinc finger protein. *Science* **260**, 530–533 (1993).
59. Searles, M. A., Lu, D. & Klug, A. The role of the central zinc fingers of transcription factor IIIA in binding to 5 S RNA. *J. Mol. Biol.* **301**, 47–60 (2000).
60. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 183–212 (2000).
61. Lai, W. S., Carballo, E., Thorn, J. M., Kennington, E. A. & Blackshear, P. J. Interactions of CCCH zinc finger proteins with mRNA. Binding of tristetraprolin-related zinc finger proteins to AU-rich elements and destabilization of mRNA. *J. Biol. Chem.* **275**, 17827–17837 (2000).
62. D'Souza, V. & Summers, M. F. Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus. *Nature* **431**, 586–590 (2004).
63. De Guzman, R. N. *et al.* Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* **279**, 384–388 (1998).
64. Subramanian, A. R. Structure and functions of ribosomal protein S1. *Prog. Nucleic Acid Res. Mol. Biol.* **28**, 101–142 (1983).
65. Bycroft, M., Hubbard, T. J., Proctor, M., Freund, S. M. & Murzin, A. G. The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell* **88**, 235–242 (1997).
66. Murzin, A. G. OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* **12**, 861–867 (1993).
67. Arcus, V. OB-fold domains: a snapshot of the evolution of sequence, structure and function. *Curr. Opin. Struct. Biol.* **12**, 794–801 (2002).
68. Schubert, M. *et al.* Structural characterization of the RNase E S1 domain and identification of its oligonucleotide-binding and dimerization interfaces. *J. Mol. Biol.* **341**, 37–54 (2004).
69. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. Structure and nucleic-acid binding of the *Drosophila* Argonaute 2 PAZ domain. *Nature* **426**, 465–469 (2003).
70. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. Nucleic acid 3′-end recognition by the Argonaute2 PAZ domain. *Nature Struct. Mol. Biol.* **11**, 576–577 (2004).
71. Yan, K. S. *et al.* Structure and conserved RNA binding of the PAZ domain. *Nature* **426**, 468–474 (2003).
72. Macrae, I. J. *et al.* Structural basis for double-stranded RNA processing by Dicer. *Science* **311**, 195–198 (2006).
73. Ma, J. B., Ye, K. & Patel, D. J. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429**, 318–322 (2004).
74. Yuan, Y. R. *et al.* Crystal structure of *A. aeolicus* argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. *Mol. Cell* **19**, 405–419 (2005).
75. Ma, J. B. *et al.* Structural basis for 5′-end-specific recognition of guide RNA by the *A. fulgidus* Piwi protein. *Nature* **434**, 666–670 (2005).
76. Song, J. J., Smith, S. K., Hannon, G. J. & Joshua-Tor, L. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305**, 1434–1437 (2004).
77. Parker, J. S., Roe, S. M. & Barford, D. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *EMBO J.* **23**, 4727–4737 (2004).
78. Parker, J. S., Roe, S. M. & Barford, D. Structural insights into mRNA recognition from a PIWI domain–siRNA guide complex. *Nature* **434**, 663–666 (2005).
79. Dominguez, C. & Allain, F. H. NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with *Bcl-x* G-tract RNA: a novel mode of RNA recognition. *Nucleic Acids Res.* **34**, 3634–3645 (2006).
80. Swanson, M. S. & Dreyfuss, G. Classification and purification of proteins of heterogeneous nuclear ribonucleoprotein particles by RNA-binding specificities. *Mol. Cell. Biol.* **8**, 2237–2241 (1988).
81. McCullough, A. J. & Berget, S. M. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.* **17**, 4562–4571 (1997).
82. Garneau, D., Revil, T., Fisette, J. F. & Chabot, B. Heterogeneous nuclear ribonucleoprotein F/H proteins modulate the alternative splicing of the apoptotic mediator Bcl-x. *J. Biol. Chem.* **280**, 22641–22650 (2005).
83. Jacks, A. *et al.* Structure of the C-terminal domain of human La protein reveals a novel RNA recognition motif coupled to a helical nuclear retention element. *Structure* **11**, 833–843 (2003).
84. Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. Modular recognition of RNA by a human Pumilio-homology domain. *Cell* **110**, 501–512 (2002). **This work beautifully illustrates how a protein can use multiple repeated structural motifs to create specific, high-affinity interactions with RNA; each domain binds a single nucleotide, but the combination of multiple domains provides exquisite specificity.**
85. Cheong, C. G. & Hall, T. M. Engineering RNA sequence specificity of Pumilio repeats. *Proc. Natl Acad. Sci. USA* **103**, 13635–13639 (2006). **Building on reference 84, this work introduces a reasonably predictive recognition code for this family of RNA-binding proteins that allows rational engineering of specificity.**
86. Worbs, M., Bourenkov, G. P., Bartunik, H. D., Huber, R. & Wahl, M. C. An extended RNA binding surface through arrayed S1 and KH domains in transcription factor NusA. *Mol. Cell* **7**, 1177–1189 (2001).
87. Shamoo, Y., Abdul-Manan, N. & Williams, K. R. Multiple RNA binding domains (RBDs) just don't add up. *Nucleic Acids Res.* **23**, 725–728 (1995). **This work examines in quantitative detail the importance of the linker in recognition of an RNA and provides a simple method for predicting the affinity of two RRMs separated by a linker of variable length.**
88. Shamoo, Y. *et al.* Both RNA-binding domains in heterogenous nuclear ribonucleoprotein A1 contribute toward single-stranded-RNA binding. *Biochemistry* **33**, 8272–8281 (1994).
89. Finger, L. D., Johansson, C., Rinaldi, B., Bouvet, P. & Feigon, J. Contributions of the RNA-binding and linker domains and RNA structure to the specificity and affinity of the nucleolin RBD12/NRE interaction. *Biochemistry* **43**, 6937–6947 (2004).

90. Lakatos, L., Szittya, G., Silhavy, D. & Burgyan, J. Molecular mechanism of RNA silencing suppression mediated by p19 protein of tombusviruses. *EMBO J.* **23**, 876–884 (2004).
91. Dunoyer, P., Lecellier, C. H., Parizotto, E. A., Himber, C. & Voinnet, O. Probing the microRNA and small interfering RNA pathways with virus-encoded suppressors of RNA silencing. *Plant Cell* **16**, 1235–1250 (2004).
92. Vargason, J. M., Szittya, G., Burgyan, J. & Tanaka Hall, T. M. Size selective recognition of siRNA by an RNA silencing suppressor. *Cell* **115**, 799–811 (2003).
93. Ye, K., Malinina, L. & Patel, D. J. Recognition of small interfering RNA by a viral suppressor of RNA silencing. *Nature* **426**, 874–878 (2003).
94. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. The structure of the flock house virus B2 protein, a viral suppressor of RNA interference, shows a novel mode of double-stranded RNA recognition. *EMBO Rep.* **6**, 1149–1155 (2005).
95. Chao, J. A. *et al.* Dual modes of RNA-silencing suppression by Flock House virus protein B2. *Nature Struct. Mol. Biol.* **12**, 952–957 (2005).
96. Ramos, A. *et al.* Role of dimerization in KH/RNA complexes: the example of Nova KH3. *Biochemistry* **41**, 4193–4201 (2002).
97. Calero, G. *et al.* Structural basis of m7GpppG binding to the nuclear cap-binding protein complex. *Nature Struct. Biol.* **9**, 912–917 (2002).
98. Buttner, K., Wenig, K. & Hopfner, K. P. Structural framework for the mechanism of archaeal exosomes in RNA processing. *Mol. Cell* **20**, 461–471 (2005).
99. Liu, Q., Greimann, J. C. & Lima, C. D. Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell* **127**, 1223–1237 (2006).
This beautiful structure provides a number of examples of protein–protein interactions between S1 and KH domains at the core of the exosome.
100. Abovich, N. & Rosbash, M. Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell* **89**, 403–412 (1997).
101. Michaud, S. & Reed, R. An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes Dev.* **5**, 2534–2546 (1991).
102. Zamore, P. D., Patton, J. G. & Green, M. R. Cloning and domain structure of the mammalian splicing factor U2AF. *Nature* **355**, 609–614 (1992).
103. Kielkopf, C. L., Lucke, S. & Green, M. R. U2AF homology motifs: protein recognition in the RRM world. *Genes Dev.* **18**, 1513–1526 (2004).
104. Andersen, C. B. *et al.* Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. *Science* **313**, 1968–1972 (2006).
105. Stroupe, M. E., Tange, T. O., Thomas, D. R., Moore, M. J. & Grigorieff, N. The three-dimensional architecture of the EJC core. *J. Mol. Biol.* **360**, 743–749 (2006).
106. Irion, U., Adams, J., Chang, C. W. & St Johnston, D. Miranda couples oskar mRNA/Staufen complexes to the bicoid mRNA localization pathway. *Dev. Biol.* **297**, 522–533 (2006).
107. Collins, R. E. & Cheng, X. Structural domains in RNAi. *FEBS Lett.* **579**, 5841–5849 (2005).
108. Li, L. & Ye, K. Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature* **443**, 302–307 (2006).
109. Reichow, S. L., Hamma, T., Ferre-D'Amare, A. R. & Varani, G. The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res.* **35**, 1452–1464 (2007).
110. Williams, B. R. PKR; a sentinel kinase for cellular stress. *Oncogene* **18**, 6112–6120 (1999).
111. Bass, B. L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817–846 (2002).
112. Nanduri, S., Rahman, F., Williams, B. R. & Qin, J. A dynamically tuned double-stranded RNA binding mechanism for the activation of antiviral kinase PKR. *EMBO J.* **19**, 5567–5574 (2000).
This work demonstrates how the second dsRBD of the dsRNA-activated kinase PKR interacts with the C-terminal kinase domain, maintaining it in an inhibited state.
113. Macbeth, M. R., Lingam, A. T. & Bass, B. L. Evidence for auto-inhibition by the N terminus of hADAR2 and activation by dsRNA binding. *RNA* **10**, 1563–1571 (2004).
114. Gelev, V. *et al.* Mapping of the auto-inhibitory interactions of protein kinase R by nuclear magnetic resonance. *J. Mol. Biol.* **364**, 352–363 (2006).
115. Li, S. *et al.* Molecular basis for PKR activation by PACT or dsRNA. *Proc. Natl Acad. Sci. USA* **103**, 10005–10010 (2006).
116. Bevilacqua, P. C. & Cech, T. R. Minor-groove recognition of double-stranded RNA by the double-stranded RNA-binding domain from the RNA-activated protein kinase PKR. *Biochemistry* **35**, 9983–9994 (1996).
117. Kim, I., Liu, C. W. & Puglisi, J. D. Specific recognition of HIV TAR RNA by the dsRNA binding domains (dsRBD1–dsRBD2) of PKR. *J. Mol. Biol.* **358**, 430–442 (2006).
118. Carpick, B. W. *et al.* Characterization of the solution complex between the interferon-induced, double-stranded RNA-activated protein kinase and HIV-I trans-activating region RNA. *J. Biol. Chem.* **272**, 9510–9516 (1997).
119. Romano, P. R. *et al.* Autophosphorylation in the activation loop is required for full kinase activity *in vivo* of human and yeast eukaryotic initiation factor 2α kinases PKR and GCN2. *Mol. Cell. Biol.* **18**, 2282–2297 (1998).
120. Zhang, F. *et al.* Binding of double-stranded RNA to protein kinase PKR is required for dimerization and promotes critical autophosphorylation events in the activation loop. *J. Biol. Chem.* **276**, 24946–24958 (2001).
121. Frazao, C. *et al.* Unravelling the dynamics of RNA degradation by ribonuclease II and its RNA-bound complex. *Nature* **443**, 110–114 (2006).
122. Antson, A. A. *et al.* Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* **401**, 235–242 (1999).
123. Oberstrass, F. C. *et al.* Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nature Struct. Mol. Biol.* **13**, 160–167 (2006).

Acknowledgements
Work in our laboratories is supported by grants from National Institutes of Health–National Institute of General Medical Sciences (G.V. and C.M.). We apologize to the many colleagues whose work could not be properly referenced owing to lack of space.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Protein Data Bank: http://www.rcsb.org/pdb/home/home.do
1C9S | 1EC6 | 1EKZ | 1M8Y | 1RGO | 1SI3 | 1U04 | 1UN6 | 1URN | 1YTU | 2ASB | 2ESE | 2FFL | 2IX1 | 2NN6
UniProtKB: http://ca.expasy.org/sprot
Nova-1 | SF1 | TIS11d

FURTHER INFORMATION
Claire Moore's homepage: http://www.tufts.edu/sackler/microbiology/faculty/moore/index.html
Gabriele Varani's homepage:
http://depts.washington.edu/varani2
Access to this links box is available online.