Published in final edited form as:
Nat Rev Mol Cell Biol. 2015 September ; 16(9): 533–544. doi:10.1038/nrm4032.

Specificity and non-specificity in RNA–protein interactions

Eckhard Jankowsky[1,2,3] and **Michael E. Harris**[2]

[1]Center for RNA Molecular Biology, School of Medicine, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106

[2]Department of Biochemistry, School of Medicine, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106

[3]Department of Physics, School of Medicine, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106

Abstract

Gene expression is regulated by complex networks of interactions between RNAs and proteins. Proteins that interact with RNA have been traditionally viewed as either specific or non-specific; specific proteins interact preferentially with defined RNA sequence or structure motifs, whereas non-specific proteins interact with RNA sites devoid of such characteristics. Recent studies indicate that the binary "specific vs. non-specific" classification is insufficient to describe the full spectrum of RNA–protein interactions. Here, we review new methods that enable quantitative measurements of protein binding to large numbers of RNA variants, and the concepts aimed as describing resulting binding spectra: affinity distributions, comprehensive binding models and free energy landscapes. We discuss how these new methodologies and associated concepts enable work towards inclusive, quantitative models for specific and non-specific RNA–protein interactions.

RNA–protein interactions are critical for the regulation of gene expression[1]. Research over the last decades has shown that RNA is invariably bound and often altered by proteins in cells, and that in biological environments RNAs generally function together with proteins as RNA–protein complexes (RNPs) [2,3]. It has also become clear that cellular RNA–protein interactions represent a very complex network, comprised of a large number of RNAs, proteins and RNA–protein interactions[4]. In addition, multitude of diseases have been linked to misregulation or malfunction of proteins that contact RNA [5–7]. Thus, deciphering RNA–protein interactions on both molecular and cellular scales is central to understanding human physiology and many diseases.

Typical eukaryotic cells contain thousands of different RNAs[8]. For every protein that interacts with RNA it is critical to understand the molecular characteristics that define whether and how the protein discriminates between different potential binding sites in these RNAs. For this purpose, proteins that interact with RNA are traditionally classified as either "specific" or "non-specific". Specific proteins associate preferentially with defined RNA sequence or structure motifs, or a combination thereof. "Non-specific" proteins associate with RNA sites that appear to be devoid of sequence or structure motifs. Roughly half of all proteins that interact with RNA proteins fall into the "non-specific" category. Examples include translation elongation and initiation factors, and proteins involved in RNA degradation[9,10]. Binding to diverse RNA sites is critical for the biological function of non-specific proteins.

Although the terms “specific” and “non-specific” are widely used, a multitude of studies that mapped RNA–protein interactions in cells or measured RNA–protein association for large numbers of sequences *in vitro*, indicate that specificity, or the lack thereof, is a considerably more nuanced problem than suggested by the binary “specific vs. non-specific” classification. As descriptions of cellular RNA–protein interaction networks move towards systems-level, quantitative models[4,11,12], and as other lines of research aim at engineering novel RNA-binding proteins[13–16], a comprehensive, quantitative view on specificity and non-specificity is required. In this review, we discuss emerging approaches aimed at this goal. To emphasize the significance of these methods and concepts for a deeper understanding of cellular RNA–protein interactions, we start with a brief overview of the tremendous complexity of RNA–protein interactions *in vivo* (in the cell). We then discuss novel methods that enable quantitative measurements of protein binding to large numbers of RNA variants, and the concepts aimed as describing resulting binding spectra: affinity distributions, binding models and free energy landscapes. Finally, we review the insights gained and the potential provided by these new methods and the associated concepts towards a nuanced, inclusive description of specific and non-nonspecific RNA protein interactions.

The complexity of RNA-protein interactions

In mammalian cells, more than 1,000 diverse proteins directly interact with RNA [1,17–19]. For the purpose of this review, we will refer to proteins that interact with RNA as RNA binding proteins (RBP), even though only a subset of these proteins function to solely bind RNA. In humans, a certain set of RBPs is expressed in all tissues investigated thus far[1]. For other RBPs, expression can vary considerably, and some are expressed exclusively in certain tissues [1,5,20,21]. Many RBPs have a modular structure, often containing multiple, different RNA interacting domains [1,22,23]. RNA interacting domains are traditionally called RNA-binding domains (RBDs), but these domains often harbor functions that exceed mere RNA binding (Tab.1). For the purpose of this review we will keep the RBD designation. The main RBD classes include enzymatic domains that chemically alter RNA (nucleotidyltransferases, ribonucleases, RNA modifying enzymes), or that couple nucleotide binding or hydrolysis to RNA binding or structural remodeling (GTPases, helicases) (Tab.1). In addition, there are numerous RBDs that only bind RNA (Tab.1). Some RBDs are found in large numbers of proteins[1,5,17]. The most frequently occurring is the R͟NA R͟ecognition M͟otif (RRM), an RNA binding module present in several hundred mammalian proteins[24]. The most common enzymatic domain is the helicase domain, found in roughly 70 human proteins that interact with RNA[17,25]. In contrast, other domains, for example RNA guanyltransferase, are found in only a single protein per organism [26]. Finally, proteins that interact with RNA vary widely in their abundance, ranging from few to 100,000 molecules per cell[27].

RNA binding is not restricted to proteins with domains that are traditionally viewed as RDBs. Recent work has revealed extensive RNA association of considerable numbers of metabolic enzymes lacking previously identified RBDs [28,18,19,29]. Other studies show association of (mostly long-non-coding) RNAs with transcription factors[30–32]. The number of proteins that demonstrably interact with RNA is thus likely to grow in the future.

The number of RNA species far exceeds the number of RBPs in typical eukaryotic cells. Human cells encode more than 20,000 different mRNAs (Fig. 1). Most cell types express between 11,000 and 15,000 at any time [33]. The diversity of mRNAs is further increased by alternative splicing [34] and by chemical modifications [35–38]. In addition to mRNAs, metazoan cells can express thousands of species of long non-coding RNAs and hundreds of micro RNAs (miRNAs), tRNAs and small nucleolar RNAs (snoRNAs) (Fig. 1). At certain stages of germ cell development, large numbers of piwi-interacting RNAs (piRNAs) are expressed[39]. On the other hand, there are only few ribosomal RNA (rRNA) and small nuclear RNA (snRNA) species (Fig. 1). Finally, cleaved RNA fragments are emerging as potentially regulatory molecules[40–42].

The various RNA types differ dramatically in their abundance. In most eukaryotic cells, rRNA accounts for roughly 80–85% of the cellular RNA mass, followed by tRNA, mRNA and snoRNAs (Fig. 1). All other RNAs together account for less than 2% of the mass (Fig. 1). At certain stages of germ cell development, these RNA mass ratios might change due to the expression of piRNAs[39]. Even within each RNA class, abundance varies widely. The expression levels for mRNAs range over four orders of magnitude[33]. A small number of expressed mRNA species often accounts for 50% of the cellular mRNA mass. For example, 50% of the mRNA mass is contributed by only 250 mRNA species (~4%) in yeast, by 900 mRNA species (~7%) in human cerebellum, and by less than 10 of mRNA species (~0.01%) in liver tissue [33]. Another factor contributing to the disparity in cellular RNA mass is that RNAs vary greatly in their length, ranging from more than 10,000 nucleotides (mRNAs and lncRNAs) to only 22 nucleotides(miRNAs) (Fig. 1).

Any given RNA is usually bound by multiple proteins [3,4]. Different proteins can either bind simultaneously, subsequently, or in a mutually exclusive manner [3,4]. Conversely, most proteins can bind multiple RNAs[43]. Some proteins, such as mRNA export factors, require the capacity to contact many diverse mRNAs[44], and the translation elongation factor Tu binds all charged tRNAs[45]. Given the number of RNAs and RBPs, the number of possible RNA–protein interactions is extremely large. Further variation is added by proteins that do not directly contact the RNA, but modulate the binding of RNA by RBPs, for example through post-translational modifications or by through interactions with RBPs [46,47]. RNAs can also interact with one another, illustrated most prominently by the interactions between miRNAs, mRNA and ceRNAs[48,49]. Given the simultaneous presence of large numbers of RNAs and RBPs and the layers of modulation of these interactions by other proteins, cellular RNA–protein interactions represent a massive set of interdependent intereactions. Most RBDs recognize sites comprised of only to 3 to 8 nucleotides and often tolerate a high degree of sequence variation in these binding sites [3]. Thus the number of potential interactions of even highly selective proteins in organisms with small transcriptomes such as yeast can be extrordinarily large.

Every interaction between an individual protein and a specific RNA site is dictated by the inherent affinity of the protein for the RNA site, the concentration of the protein, the concentration of the RNA, the competition from other RNAs for association with the protein, and the competition of other proteins for the RNA binding site. In addition, proteins that interact with or modify RBPs can profoundly impact RNA binding patterns. Therefore, it is not surprising that substrate selection by a given protein rarely conforms to a binary specific vs. non-specific model. Yet, the challenge remains to devise models that describe RNA protein interactions in sufficient quantitative detail to allow predictions of the RNA binding pattern of individual proteins under a defined set of parameters. A critical first step towards this goal is addressed by approaches that quantitatively assess binding of proteins to many different RNA sites.

Measuring protein binding to many RNAs

Several methods have been developed to determine protein binding sites on RNAs on a transcriptome-wide scale[50,51]. The techniques rely on covalent crosslinking of protein to RNA by UV irradiation (CLIP and derivatives)[52–55], or on immunoprecipitation of RNA-bound proteins with a chemical crosslinker (RiPIT [56]) or without [57]. The crosslinked RNA fragments are identified by next generation sequencing or microarray analysis. These methods represent a quantum leap forward with respect to visualizing protein binding patterns on RNAs, often revealing binding to numerous different sites on large numbers of RNAs. The binding sites often allow the definition of consensus motifs for protein binding [43]. Although powerful and highly instructive, these techniques do not currently provide quantitative data necessary to assess affinity or binding and dissociation kinetics of RNA–protein interactions.

Other, novel approaches aim at quantitatively measuring protein binding to large numbers of RNA variants *in vitro*. Recently, *in vitro* selection (SELEX) was combined with high throughput sequencing [13,58,59]. Traditional SELEX combines multiple rounds of selection and amplification of has been traditionally used to identify RNA species preferentially bound by RBPs. The combination of SELEX with next generation sequencing allows the analysis of a much larger number of sequences than obtained with classic SELEX, and thus enables insight into RNA binding by RBP beyond the tightest bound species [13,58,59]. SELEX has also been employed to determine binding preferences of RBPs in the cell [60]. However, SELEX approaches bias the analysis towards the highest-affinity targets, even when combined with next generation sequencing.

To circumvent this bias towards the highest-affinity targets techniques have been developed that directly analyze interactions of proteins with large populations of diverse RNAs (Box 1). These methods bypass the selection and amplification cycles of the SELEX procedure, and allow measurements of both weakly and tightly bound RNA species. Some of these techniques are analogous to high throughput methods for investigating the binding of transcription factors to large numbers of DNA sequences[61]. All of the approaches measure differences in protein binding to a pool of diverse RNA substrates that contain regions of randomized nucleotides (Box 1).

Box 1

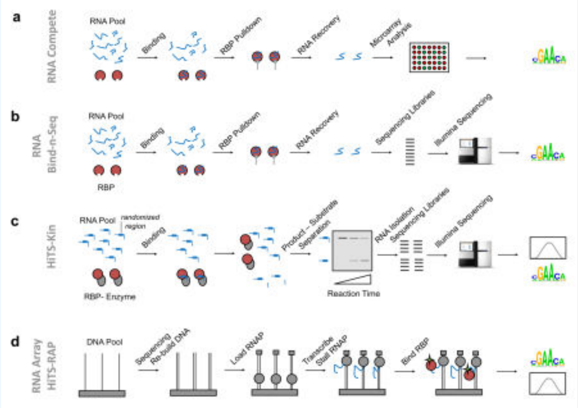
Techniques for measuring protein binding to many RNA sequences *in vitro*

a. RNA Compete. A pool of RNA species that contain a region of randomized sequence is incubated *in vitro* with a specific RNA binding protein (RBP) [129]. The RBP is pulled-down, the bound RNAs are recovered and their sequence is determined by microarrays. The method has been used to determine sequence motifs for RNAs that bind tightest to a given protein [129,130].

b. RNA Bind-n-Seq (RBNS). This approach is similar to RNA Compete, however, the bound RNAs are analyzed by next generation sequencing [103]. The number of sequences that can be measured simultaneously is limited by the throughput of the sequencer. This limit is currently at approximately (2.5 – 5) × 10^8 RNAs, which corresponds to a sequence space of 9 – 10 randomized nucleotides. RBNS has been used to obtain sequence motifs for RNA variants with the highest equilibrium binding affinity to a given RBP [103].

c. High Throughput Sequencing Kinetics (HiTS-Kin). This approach follows the enzymatic processing of various RNA substrates in a reaction that depends on an RBP, thus measuring functional binding of an RNA by the protein [67]. Processed and non-processed RNA species are separated by gel electrophoresis, although other separation techniques can be employed. The ratios of processed or non-processed RNAs, or both, as a function of time are analyzed by next generation sequencing. HiTS-Kin provides association and dissociation kinetics and has been used to determine functional affinity distributions [67]. The number of sequences that can be monitored simultaneously by HiTS-Kin is limited by the sequencing capacity and the desired quantitative coverage of the sample. The approach can be readily adapted to different experimental systems and to reactions *in vivo*, provided that reactive and unreactive RNA species can be separated.

d. RNA array and High-Throughput Sequencing-RNA Affinity Profiling (HiTS-RAP). These techniques utilize a modified Illumina next generation sequencing protocol to directly visualize the RNA–protein interactions in the sequencer flowcell[90,131]. A pool of DNA sequences with randomized regions is immobilized in a sequences flowcell. The respective sequences are identified by a round of sequencing. Subsequently, each DNA serves as template for RNA polymerase to transcribe an RNA from each DNA template. Following transcription, the polymerase is stalled (in RNA array by biotin-streptavidin at the terminus of the DNA helix; in HiTS-RAP by binding of the protein Tus to a *Ter* site in the DNA). The transcribed RNA remains bound to the stalled polymerase, thus allowing identification of each RNA species. The RNA binding protein is fluorescently labeled. The interaction of the protein with the RNA is directly monitored by measuring fluorescence changes at the positions that correspond to the different immobilized RNAs. Proteins can be flowed in and out the flowcell multiple times and at different concentrations, providing readout of binding and dissociation kinetics in real time[90]. Measuring the increase in fluorescence over time for each individual RNA species provides association rate constants. Dissociation rate constants are measured by challenging RNA-bound protein with buffer and monitoring the decrease in fluorescence over time. The RNA array and HiTS-RAP approaches are conceptually similar to techniques for measuring kinetics of protein-RNA interactions by single molecule fluorescence via total internal reflection [132].



RNA–protein affinity distributions

Most studies that map RNA–protein interaction on a transcriptomic scale show that RBPs often bind to RNA sites that vary considerably in sequence or structure [43,62]. This is expected for proteins considered “non-specific” binders, but appears to contradict the notion of “sequence specific” binding proteins. Similar observations have been made for DNA binding by “specific” transcription factors [63]: *In vitro* measurements of intrinsic affinities of transcription factors for all possible sequence variants of DNA oligomers showed that each protein had a wide range of binding affinities to the different sequence variants[64–66]. Differences between low and high affinity sites are often considerable, spanning several orders of magnitude with respect to equilibrium dissociation constants[63–66]. To describe the entire spectrum of affinities seen for a given DNA or RNA binding protein towards all possible DNA or RNA species, it is useful to employ affinity distributions [67], histogram plots of substrate variants with similar affinities (Fig. 2). Affinity distributions have revealed incremental contributions of the nucleotides in the binding site to the equilibrium binding free energy, rather than a drastic difference between nucleotide composition in preferred and non-preferred sites (Fig. 2). For “sequence specific” transcription factors, physiologically preferred binding sites cluster at the high-affinity region of the distribution[61,67,68].

A complete RNA affinity distribution, in contrast to measurements of only sequences with high protein-affinity, has so far only been reported for C5, the protein subunit of RNaseP from *Escherichia. coli.* [67] (Fig. 2). The shape of the measured affinity distribution was highly similar to those seen for transcription factors[67], also suggesting incremental contributions of the nucleotides in the binding site to the binding free energy. Incremental differences between sequence variants explain why proteins can bind with similar affinity to a range of seemingly divergent sequences (Fig. 2), which is particularly significant for RNA binding proteins, because cognate sites for most RBDs encompass only 3–8 nucleotides[3]. Potential binding sites of this size occur with few substitutions at very high frequency even in small genomes. Ambivalence regarding protein binding preferences might be amplified by the varying expression levels of the RNAs[33]. At limiting concentrations of protein, low affinity non-consensus sites in highly expressed RNAs can efficiently compete for protein binding with high affinity consensus sites in an RNA expressed at a lower level. This scenario might be one of the reasons why in the cell proteins that are considered specific are often found bound to sites with relatively poor match to their consensus motifs[62]. It is an open question whether or not binding to degenerate sites has biological consequences other than protein sequestration.

Affinity distributions also provide the means to comprehensively quantify the specificity of a given RBP or RBD, although to our knowledge, this has not yet been reported. Most of the available affinity distributions for DNA and RNA binding proteins appear to follow a Gaussian distribution for binding free energies (Fig. 2b). The width of the distribution thus provides an objective parameter for the capacity of an RBP and RBD to globally discriminate between RNA substrate variants. Multi-modal distributions are also conceivable, these would describe different binding modes, which could arise, for example, through formation of stable RNA structures by a subset of substrates.

The structural basis for affinity distributions are not currently understood. However, a recent pioneering study investigated the structural basis for a range of affinities that the bacterial RBP RsmE shows towards different substrate variants [69], although no complete affinity distribution was measured. For RsmE, conformational adaptation of protein side-chains and RNA are responsible the range of affinities [69].

Binding models

As noted, affinity distributions are useful because they represent a non-biased description of protein binding to unstructured RNA, to a defined RNA structure, or to a combination of both. For proteins that bind to unstructured RNA, sequence variants in the high affinity region of the distribution share a consensus sequence motif [67], which can be expressed as a sequence probability logo [61] (Fig. 2b). Other regions of the distribution do not share a sequence consensus (Fig. 2b). Consensus sequences describe the probability by which a given nucleotide is present at a given position in the binding site for a subset of all sequence variants[61,68]. A larger number of sequence variants in a given subset of the distribution decreases the probabilities and thus reduces the strength of the consensus. There are several approaches to delineate consensus motifs from binding site-data, obtained either *in vitro* or *in vivo*[70–77]. A consensus motif can guide a qualitative prediction of whether or not a protein binds well to a certain motif. However, consensus motifs are not binding models, because a consensus motif does not allow affinity calculations for different sequence variants, nor does it provide information on the characteristics of the entire affinity distribution[61,68].

The simplest model to describe binding of a protein to all RNA sequence variants is the Position Weight Matrix (PWM)[61,68]. A PWM is a score calculated for each nucleotide at each position in the binding site (Fig. 3a). Thie sum of the individual nucleotide scores for a given sequence provides a score for this sequence variant [61]. The PWM can also be visualized as logo[68]. It is important to note that a PWM logo thus differs from a probability logo, discussed above. If affinities are expressed as binding free energies, a PWM becomes an energy score, describing the energetic contribution of each nucleotide at each position to the binding free energy [61,68]. A PWM implies that the nucleotides at each position contribute independently of each other to the binding of the protein[61,68,78]. In reality, the impact of a nucleotide at one position is often influenced by the surrounding sequence. PWMs often explain only a subset of the observed experimental variance in affinities[61,66,68,79], and frequently fail to accurately explain the highest and the lowest observed affinities [67].

A better explanation of the observed experimental variance is usually accomplished by considering the coupling of contributions from different nucleotide positions in the binding model[67,80,81] (Fig. 3b). Couplings are considered by assigning a score for each combination of nucleotides and then summing up the score for the combinations resulting in a given sequence [67]. Even the incorporation of a modest number of pairwise couplings (called either Pairwise Interaction Matrix, PIM, or Dinucleotide Weight Matrix, DWM) often improves the binding model considerably[67,80,81]. However, it is critical to carefully evaluate that an improved fit does not simply result from the incorporation of more variables in the model. Of note, only roughly 20–30 % of the entire sequence space is needed in order to produce an unambiguous binding model, provided the sequences cover the entire range of the affinity distribution [67]. Interdependencies between neighboring nucleotides in binding sites of DNA interacting proteins have also been described by Hidden Markov models [82,83], and these models are applicable to RBPs as well.

While the consideration of interdependencies between two nucleotides generally improves the binding model, further improvements can be accomplished by considering higher order couplings between more than two nucleotides in the binding model[78]. An array of approaches to accomplish this goal have been developed for DNA binding proteins, including higher order Hidden Markov models[84], neural network analysis[85], decision tree guided approaches[86], higher order Bayesian networks[87], and approaches that incorporate structural information about the protein[88]. Neural network analysis has been applied to RNA–protein interactions measured *in vitro* with the HiTS-Kin approach[67]. In this case, the neural network analysis considering higher order couplings between multiple nucleotides did not markedly improve the fit of the model to the data for the C5 protein, suggesting that pairwise couplings between mostly adjacent nucleotides are the major contributor to protein binding in the tested case[67]. Hidden Markov models were also used to improve rules predicting RNA binding patterns for the splicing factor PTB *in vivo*[89].

Free energy landscapes

Substrate affinities for proteins that interact with RNA usually refer to equilibrium binding constants, which express the energetic difference between ground state (protein and RNA are unbound) and product state (protein and RNA are bound) in a one-step binding reaction (Box 2, part **a**). However, differences in equilibrium binding affinity between substrate variants can arise from alterations in ground, transition, or product state energies, or from combinations thereof (Box 2, part **a**). These alterations can only be assessed through measurements of association and dissociation rate constants for the substrate variants. To date, only few studies have reported rate constants for many substrates[67,90], and to our knowledge only one (REF.[90]) reported both association and dissociation rate constants — for the binding of the MS2 coat protein to a large set of variants of the cognate RNA hairpin[90]. In this study, differences in substrate preferences were mainly due to variations in substrate association rate constants, with comparably small contributions by dissociation rate constants. These observations suggest that for the MS2–substrate system differences in RNA binding are mainly due to variations in ground state energies, most likely reflecting the significance of RNA structure for substrate binding by MS2 [90].

Box 2

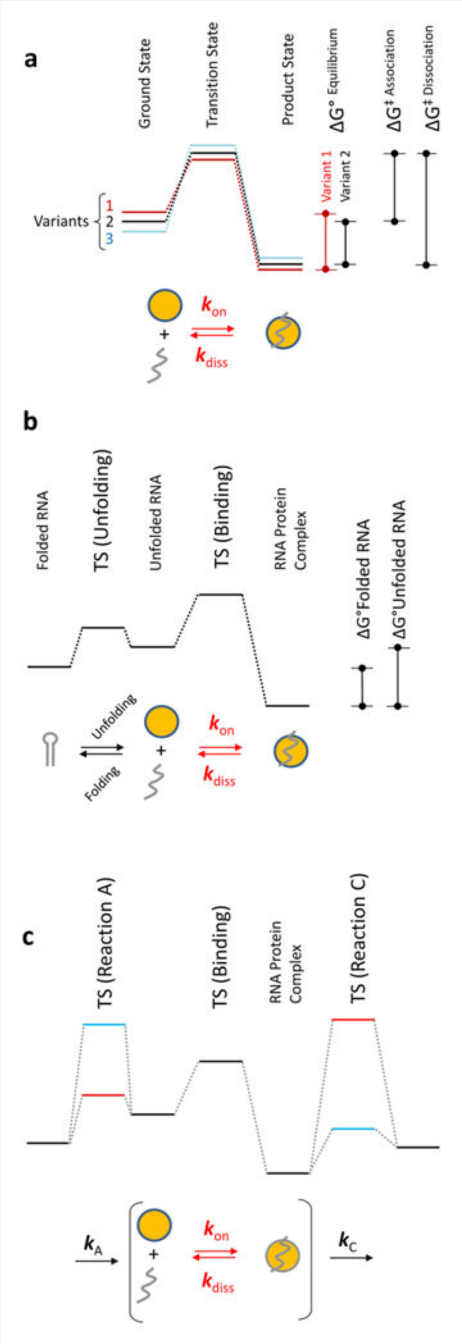
Free energy landscapes for RNA–protein interactions

a. Free energy landscape of a single step of a reversible binding reaction between a protein and three RNA variants. The different binding affinities of sequence variants are reflected in different equilibrium free energy changes ($\Delta G^{\circ \text{Equilibrium}}$). Different equilibrium binding affinities can result from differences in ground-, transition- or product-state free energies, or through combinations thereof, which correspond to changes in activation free energy for association ($\Delta G^{\ddagger \text{Association}}$), dissociation ($\Delta G^{\ddagger \text{Dissociation}}$), or both.

b. Free energy landscape of a binding reaction between a protein and structured RNA. Only one RNA variant is shown for simplicity. In this example the RNA binds in the unfolded state, however the mechanism is general and also applies to structural transitions more complex than hairpin unfolding. The unfolding step affects the equilibrium free energy change (ΔG°), and thus the binding affinity. Depending on the transition state (TS) energy for the unfolding step, RNA structures can also impact the association and possibly dissociation kinetics.

c. Kinetic context of an RNA–protein binding reaction. Only one RNA variant is shown for simplicity.

The scheme shows ground and transition states (TS) for three consecutive reaction steps. The protein-RNA binding step is the middle step. Intrinsic specificity can only translate into biological specificity in the context of the reaction shown for the scenario indicated by the red colored transitions state energies for reaction A (rate constant k_A) and C (rate constant k_C). All other combinations of transition state energies reduce or virtually eliminate intrinsic specificity, as reflected in the isolated binding reaction.



While comparable data for other RNA–protein interactions have not yet been reported, RNA structure is likely to impact selectivity even for proteins that bind to presumably unstructured sites (Box 2, part **b**). It is expected that a subset of a randomized substrate population forms at least transient secondary structures [91,92]. The unfolding of even quite unstable structures will affect the substrate's ground state and thereby impact the overall affinity distribution. Although it is known that sequestration of protein binding sites by RNA structures can greatly impact protein binding *in vitro*[93] and *in vivo*[94], it has not been explored to which degree more subtle changes in substrate ground state free energies contribute to binding specificity. The potential impact of even transient RNA structures on substrate specificity emphasizes the role of the RNA sequence surrounding the binding site for RNA–protein interactions.

Specific vs. non-specific interactions

As noted above, affinity distributions of RBPs measured *in vitro* and RNA bindings patterns of numerous RBPs measured in cells have raised questions regarding the widely used classification of RBPs as specific vs. non-specific. However, a more nuanced, quantitative view on specificity and non-specificity is emerging, based on the recent technical advances (Box 1), and the conceptual approaches discussed above: affinity distributions, binding models, and free energy landscapes

“Specific” vs. “non-specific” RBPs

The vast majority of studies on RNA–protein interactions have focused on “specific” RBPs, even though non-specific proteins are numerous and perform many important biological functions. A recent study determined the affinity distribution for a non-specific *E.* coli protein, the C5 subunit of RNase P [67]. C5 binds, in conjunction with the catalytic RNA unit of RNaseP, to all cellular tRNA precursors at a completely degenerate binding site [67,95,96]. Despite the lack of a consensus binding motif in its physiological substrates, the affinity distribution for C5 was highly similar to those seen for highly specific proteins [67]. Similarly to specific RBPs, the high-affinity region of the distribution for C5 showed a consensus sequence, indicating that C5 is inherently specific towards certain sequences. In contrast to specific RBPs, the physiological substrates for C5 do not fall in the high affinity region of the distribution, but are found in the median range of the distribution, which does not have a consensus, and where large differences in sequence have only small effects on affinity (Fig. 2). Notably, defined binding models could be readily derived from the C5 affinity distribution [67]. The data suggest that the differences between specific and non-specific RBPs are not inherent features of proteins. Rather, specific and non-specific binding modes represent different parts of the affinity distribution (Fig. 2b).

Intrinsic specificity even for “non-specific” RBPs is perhaps not surprising, given that protein and RNA surfaces on the binding interface have irregular features. Some RNA species are thus more likely to form favorable interactions with a protein than others. This notion probably applies to the vast majority of RNA–protein interactions. A possible exception are proteins that bind exclusively to the backbone of an A-form RNA helix, because the backbone of an A-form RNA helix is thought to be structurally similar for diverse sequences[97,98]. Yet, helices dynamically open and close locally in a sequence-dependent manner[97,99], and may be distorted upon protein binding, as seen for dsRDB–RNA complexes[100].

The kinetic context for RNA–protein interactions

The RNA binding study of C5 also highlighted the significance of the context in which a binding reaction occurs. One critical and perhaps obvious aspect for this context is the availability of substrates in the transcriptome. For C5, most of the tightest binding sequence variants are not present in the expressed substrates. RNA structure also plays an important role for the context of a binding reaction, as discussed above. Even transient sequestration of a protein binding site affects thermodynamic stability of the RNA-protein complex (Box 2, part **b**), and thus RNA sequences outside the immediate protein binding site can impact protein binding.

A third, potentially highly significant factor is the kinetic context — the kinetics of the reactions that precede and follow the binding step (Box 2, part **c**). This kinetic context is dictated by the concentration of the protein, by the concentration of the RNA, by the rate constants for substrate binding and dissociation, and how these rate constants compare to rate constants of the steps that precede and follow the binding step (Box 2, part **c**). The intrinsic specificity of the protein for any given RNA substrate variant is given by the ratio of rate constants for substrate binding and dissociation(Box 2, part **a**). However, intrinsic specificity translates only into near maximal specificity if the step preceding the binding is fast compared to the binding step, and the step following the binding step is slow compared to both binding and dissociation. All other scenarios neutralize intrinsic specificity to various degrees (Box 2, part **c**). Therefore, an inherently highly specific protein can be readily operated under an entirely non-specific regime, or a protein can be toggled between non-specific to specific modes, solely through changes in rate constants of steps unrelated to binding, or through changes in RNA or protein concentrations. The kinetic context is likely to be affected, dictated, and modulated by proteins that may or may not directly interact with the RBP in question. While we are not aware of studies that have explicitly tested kinetic context for RBPs, this context is likely to contribute to the sometimes wide range of binding sites seen during the transcriptome-wide mapping of RNA binding sites for many proteins.

Given the significance of kinetic context for substrate preference in processes with many steps, we believe it is useful to distinguish between the observed, biological specificity, and the intrinsic specificity of a protein towards substrate variants. The biological specificity is the preference of a protein for sequence variants *in vivo*, revealed by techniques like CLIP. The intrinsic specificity is the preference of a protein for sequence variants when only the binding reaction is examined *in vitro* and is reflected in the affinity distribution. Intrinsic specificity is equivalent to the classic definition of “specificity” for enzymatic reactions: $F^{\text{Specificity}} = (k_{\text{cat}}/K_{\text{m}})^{\text{Substrate1}}/(k_{\text{cat}}/K_{\text{m}})^{\text{Substrate2}}$ (REF[101]). An obvious challenge is to quantitatively define the connection between intrinsic specificity and biological specificity for RBPs. The first attempts in this direction have been reported, integrating *in vivo* and *in vitro* specificity measurements[102,103]. Even without advanced quantitative models to bridge intrinsic and biological specificity, the combination of measurements on both levels enhances the accuracy of predictions of protein binding sites in the cell.

Strategies to increase or decrease intrinsic specificity of RBPs

For many proteins the intrinsic specificity of their individual RBDs appears to be insufficient to accomplish the biologically required binding accuracy for RBPs [3,22,23,104]. Strategies have therefore evolved to enhance the intrinsic specificity of many proteins in order to better discriminate between cognate and non-cognate binding sites. On the other hand, proteins that need to interact with diverse RNA sites in an indiscriminate fashion must employ strategies to compensate for unavoidable intrinsic specificities of their RBDs.

Intrinsic specificity can be enhanced by increasing the RNA binding site-size of the RBD, in order to recognize more nucleotides (Fig. 4a,b). However, there can be a trade-off between increase in binding free energy and the number of alternative binding sites with similar free energies. A larger binding site is expected to bind the target variant tighter than a small site, but a larger binding site can also bind non-target variants tighter, and discrimination between target and non-target sites does not neccessarily increase (Fig. 4a)[105]. However, discrimination between target and non-target sites depends on whether or not additional nucleotides contribute independently to overall affinity. Independent contributions of nucleotides result in only modest increases of discrimination with increasing bind site size (Fig. 4a). In contrast, energetic coupling between certain nucleotides can result in large effects with binding site size increase (Fig. 4b). An increase in binding site size that is thought to lead to enhanced specificity is seen for the RRM[24,106], although so far no comprehensive binding analysis (affinity distribution) was reported for an RRM. However, it was clearly demonstrated that changes in binding site size are accomplished through alternative RNA binding modes on the core RRM fold and by involvement of loops in the RRM–RNA interaction [24].

A widely observed strategy to affect intrinsic specificity of RBPs is the inclusion of multiple RBDs in a single protein (Fig. 4c). As noted, a large fraction of the proteins that interact with RNA contain multiple, often different RBDs [1,22,23]. This modular architecture results in proteins with affinity distributions that are combination of the affinity distributions of their individual RBDs (Fig. 4). These combinations can enhance binding specificity (i) if the affinity distributions of the different RBDs favor similar sequence variants, or (ii) if they favor different sequence variants in a non-compensatory fashion (Fig. 4). A modular protein architecture can also enable proteins to recognize non-contiguous sequences [23] and thus intervening RNA sequences can become important contributors to specificity[107–109]. In addition, protein regions that link different RDBs can further modulate the contribution of each RBD to the protein's overall RNA binding and specificity, and even promote cooperativity between RBDs [110]. Multiple RBDs with different inherent affinity distributions can also compensate for each other in a given protein and lead to uniform binding of an RBP to a wide range of diverse substrates (Fig. 4). This is seen for the translation elongation factor Tu (EF-Tu), which binds to all charged tRNAs with similar affinity [111]. EF-Tu contains a binding site for the tRNA and one for recognizing the amino acid [112]. The binding energies for tRNAs and amino acids at each site differ, but compensate for their respective differences, thereby resulting in nearly uniform binding for all correctly charged tRNAs[111].

Multiple RBDs do not necessarily need to be part of the same protein, but can be encoded by different yet interacting proteins (Fig. 4c). This is widely seen in RNA-protein complexes [113–115], including in large RNA–protein assemblies such as the spliceosome[116–120] or the eukaryotic translation initiation machinery [121–123]. Moreover, multiple modular RBDs can assemble on the same RNA substrate, further increasing selectivity [23]. An advantage of combining different proteins for binding to given RNA site is the possibility of regulating their interactions through variations in concentration and post-translational modifications of the individual proteins[46]. Different proteins can bind cooperatively or anti-cooperatively, and these modes of protein–protein interactions can further amplify intrinsic specificity, or provide compensation for the intrinsic specificity of individual RBDs. Multiple identical RBDs can also assemble in homo-oligomers of RNA-interacting proteins [69], and can thereby enhance selectivity for longer target sites [109,124].

Future perspective

High throughput sequencing methods have opened new possibilities to measure and understand specificity and non-specificity in RNA–protein interactions, both *in vitro* and *in vivo*. It is now possible to directly determine affinities for all, or at least for a large number of possible binding site-sequence variants for most RBDs *in vitro*, and to derive comprehensive binding models. These new tools have already provided important insight into principles that underlie binding specificity and non-specificity. Although not all of these techniques can yet be readily applied in every laboratory, it is likely that binding models will emerge for many more RBDs and RBPs over the next years.

Future challenges include the integration of quantitative binding models with structural data. To date, most structures of RBPs are solved with only a single RNA substrate, usually representing a high affinity target; only in very few cases structures exist for low affinity targets [69] or alternative substrates [125]. Yet these data, combined with comprehensive binding models will be most instructive for linking structure and intrinsic specificity [125]. It will be equally important to determine binding models for proteins with mutated RBDs and, if possible, to integrate structural and binding models for such mutant proteins. Comparisons of binding models for wild type and mutated proteins might also be an inroad to dissect the virtually unexplored impact of transient RNA structure and kinetic context on RNA–protein interactions in the cell.

Ultimately, we wish to devise models that accurately describe and possibly predict the RNA binding patterns for proteins *in vivo*. This will require quantitative models that integrate RNA binding *in vitro* and *in vivo* with other aspects of RNA biology. An important step towards such models has been recently made using techniques that assess RNA secondary structures *in vivo* [126–128]. A critical, yet unconquered barrier for the development of quantitative models of RNA–protein interactions is the lack of methods to determine kinetics of RNA–protein binding *in vivo* for individual RNA sites.

Acknowledgments

We thank the members of the Jankowsky and the Harris groups for discussions on the subject of this review. Work in the author's laboratories is supported by NIGMS.

References

1. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nature Rev Gen. 2014; 15:829–845.
2. Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. Science. 2005; 309:1514–1518. [PubMed: 16141059]

3. Mitchell SF, Parker R. Principles and Properties of Eukaryotic mRNPs. Mol Cell. 2014; 54:547–558. [PubMed: 24856220]

4. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. Nature Rev Gen. 2010; 11:75–87.

5. Gerstberger, S.; Hafner, M.; Ascano, M.; Tuschl, T. Systems Biology of RNA Binding Proteins. In: Yeo, GW., editor. Exp Med Biol. Springer; 2014. p. 1-55.

6. Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. Trends Genet. 2013; 29:318–327. [PubMed: 23415593]

7. Scheper GC, van der Knaap MS, Proud CG. Translation matters: protein synthesis defects in inherited disease. Nat Rev Genet. 2007; 8:711–723. [PubMed: 17680008]

8. McGettigan PA. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol. 2013; 17:4–11. [PubMed: 23290152]

9. Parker R, Song H. The enzymes and control of eukaryotic mRNA turnover. Nature Struct Mol Biol. 2004:121–127. [PubMed: 14749774]

10. Aitken CE, Lorsch JR. A mechanistic overview of translation initiation in eukaryotes. Nature Struct Mol Biol. 2012; 19:568–576. [PubMed: 22664984]

11. Janga SC, Mittal N. Construction, structure and dynamics of post-transcriptional regulatory network directed by RNA-binding proteins. Adv Exp Med Biol. 2011; 722:103–107. [PubMed: 21915785]

12. Gerstberger S, Hafner M, Tuschl T. Learning the language of post-transcriptional gene regulation. Genome Biol. 2013; 14:130. [PubMed: 23998708]

13. Campbell ZT, Valley CT, Wickens M. A protein-RNA specificity code enables targeted activation of an endogenous human transcript. Nature Struct Mol Biol. 2014; 21:732–738. [PubMed: 24997599]

14. Wang Y, Wang Z, Tanaka Hall TM. Engineered proteins with Pumilio/fem-3 mRNA binding factor scaffold to manipulate RNA metabolism. FEBS J. 2013; 280:3755–3767. [PubMed: 23731364]

15. Chen Y, Varani G. Engineering RNA-binding proteins for biology. FEBS J. 2013; 280:3734–3754. [PubMed: 23742071]

16. Choudhury R, Tsai YS, Dominguez D, Wang Y, Wang Z. Engineering RNA endonucleases with customized sequence specificities. Nature Commun. 2012; 3:1147. [PubMed: 23093184]

17. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. 2002; 30:1427–1464. [PubMed: 11917006]

18. Baltz AG, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol Cell. 2012; 46:674–690. [PubMed: 22681889]

19. Castello A, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 2012; 149:1393–1406. [PubMed: 22658674]

20. Thomson T, Lin H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annu Rev Cell Dev Biol. 2009; 25:355–376. [PubMed: 19575643]

21. Brook M, Smith JW, Gray NK. The DAZL and PABP families: RNA-binding proteins with interrelated roles in translational control in oocytes. Reproduction. 2009; 137:595–617. [PubMed: 19225045]

22. Singh R, Valcárcel J. Building specificity with nonspecific RNA-binding proteins. Nat Struct Mol Biol. 2005; 12:645–653. [PubMed: 16077728]

23. Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. Nature Rev Mol Cell Biol. 2007; 8:479–490. [PubMed: 17473849]

24. Cléry A, Blatter M, Allain FH. RNA recognition motifs: boring? Not quite. Curr Opin Struct Biol. 2008; 18:290–298. [PubMed: 18515081]

25. Fairman-Williams ME, Guenther U-P, Jankowsky E. SF1 and SF2 helicases: family matters. Curr Opin Struct Biol. 2010; 20:313–324.10.1016/j.sbi.2010.03.011 [PubMed: 20456941]

26. Ghosh A, Lima CD. Enzymology of RNA cap synthesis. Wiley Interdiscip Rev RNA. 2010; 1:152–172. [PubMed: 21956912]

27. Firczuk H, et al. An in vivo control map for the eukaryotic mRNA translation machinery. Mol Syst Biol. 2013; 9:635. [PubMed: 23340841]
28. Hentze MW, Preiss T. The REM phase of gene regulation. Trends Biochem Sci. 2010; 35:423–426. [PubMed: 20554447]
29. Mitchell SF, Jain S, She M, Parker R. Global analysis of yeast mRNPs. Nature Struct Mol Biol. 2013; 20:127–133. [PubMed: 23222640]
30. Ng SY, Bogu GK, Soh BS, Stanton LW. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. Mol Cell. 2013; 51:349–359. [PubMed: 23932716]
31. Di Ruscio A, et al. DNMT1-interacting RNAs block gene-specific DNA methylation. Nature. 2013; 503:371–376. [PubMed: 24107992]
32. Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. Nat Rev Mol Cell Biol. 2014; 15:749–760. [PubMed: 25269475]
33. Ramsköld D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol. 2009; 5:e1000598. [PubMed: 20011106]
34. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010; 463:457–463. [PubMed: 20110989]
35. Liu N, et al. Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. RNA. 2013; 19:1848–1856. [PubMed: 24141618]
36. Dominissini D, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012; 485:201–206. [PubMed: 22575960]
37. Pan T. N6-methyl-adenosine modification in messenger and long non-coding RNA. Trends Biochem Sci. 2013; 38:204–209. [PubMed: 23337769]
38. Carlile TM, et al. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature. 2014; 515:143–146. [PubMed: 25192136]
39. Weick EM, Miska EA. piRNAs: from biogenesis to function. Development. 2014; 141:3458–3471. [PubMed: 25183868]
40. Thompson DM, Lu C, Green PJ, Parker R. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA. 2008; 14:2095–2103. [PubMed: 18719243]
41. Ivanov P, Emara MM, Villen J, Gygi SP, Anderson P. Angiogenin-induced tRNA fragments inhibit translation initiation. MOl Cell. 2011; 43:613–623. [PubMed: 21855800]
42. Saikia M, et al. Angiogenin-cleaved tRNA halves interact with cytochrome c, protecting cells from apoptosis during osmotic stress. Mol Cell Biol. 2014; 34:2450–2463. [PubMed: 24752898]
43. Milek M, Wyler E, Landthaler M. Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. Semin Cell Dev Biol. 2012; 23:206–212. [PubMed: 22212136]
44. Müller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat Rev Genet. 2013; 14:275–287. [PubMed: 23478349]
45. Agirrezabala X, Frank J. Elongation in translation as a dynamic interaction among the ribosome, tRNA, and elongation factors EF-G and EF-Tu. Q Rev Biophys. 2009; 42:159–200. [PubMed: 20025795]
46. Jangi M, Sharp PA. Building Robust Transcriptomes with Master Splicing Factors. Cell. 2014; 159:487–498. [PubMed: 25417102]
47. Zhou Z, Fu XD. Regulation of splicing by SR proteins and SR protein-specific kinases. Chromosoma. 2013; 122:191–207. [PubMed: 23525660]
48. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136:215–233. [PubMed: 19167326]
49. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014; 505:344–352. [PubMed: 24429633]
50. König J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2012; 13:77–83. [PubMed: 22251872]
51. McHugh CA, Russell P, Guttman M. Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol. 2014; 15:203. [PubMed: 24467948]

52. Ule J, et al. CLIP identifies Nova-regulated RNA networks in the brain. Science. 2003; 14:1212–1215. [PubMed: 14615540]

53. Licatalosi DD, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature. 2008; 456:464–469. [PubMed: 18978773]

54. Hafner M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010; 141:129–141. [PubMed: 20371350]

55. König J, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol. 2010; 17:909–915. [PubMed: 20601959]

56. Singh G, Ricci EP, Moore MJ. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. Methods. 2014; 65:320–332. [PubMed: 24096052]

57. Singh G, et al. The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell. 2012; 151:750–764. [PubMed: 23084401]

58. Campbell ZT, et al. Cooperativity in RNA-Protein Interactions: Global Analysis of RNA Binding Specificity. Cell Rep. 2012; 1:570–581. This study combines *in vitro* selection, high-throughput sequencing, and sequence distributions. [PubMed: 22708079]

59. Ozer A, Pagano JM, Lis JT. New Technologies Provide Quantum Changes in the Scale, Speed, and Success of SELEX Methods and Aptamer Characterization. Mol Ther Nucleic Acids. 2014; 5:e183. [PubMed: 25093707]

60. Lorenz C, et al. Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts. Nucleic Acids Res. 2010; 38:3794–3808. [PubMed: 20348540]

61. Stormo G, Zhao Y. Determining the specificity of protein - DNA interactions. Nat Rev Genetics. 2010; 11:751–760. [PubMed: 20877328]

62. Sanford JR, et al. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res. 2009; 19:381–394. [PubMed: 19116412]

63. Badis G, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009; 324:1720–1723. [PubMed: 19443739]

64. Rowe W, et al. Analysis of a complete DNA-protein affinity landscape. J R Soc Interface. 2009; 7:397–408. [PubMed: 19625306]

65. Nutiu R, et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat Biotechnol. 2011; 29:659–664. [PubMed: 21706015]

66. Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. Science. 2007; 315:233–237. [PubMed: 17218526]

67. Guenther UP, et al. Hidden specificity in an apparently nonspecific RNA-binding protein. Nature. 2013; 502:385–388. The paper introduces the HiTS-Kin method and measures affinity distributions for an RBP without canonical specificity. [PubMed: 24056935]

68. Stormo GD. Modeling the specificity of protein-DNA interactions. Quant Biol. 2013; 1:115–130. [PubMed: 25045190]

69. Duss O, Michel E, Diarra dit Konté N, Schubert M, Allain FH. Molecular basis for the wide range of affinity found in Csr/Rsm protein-RNA recognition. Nucleic Acids Res. 2014; 42:5332–5346. This study investigates differences in structures for high and low-affinity targets of an RBP. [PubMed: 24561806]

70. Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol. 2014; 15:R17. [PubMed: 24451197]

71. Pancaldi V, Bähler J. In silico characterization and prediction of global protein-mRNA interactions in yeast. Nucleic Acids Res. 2011; 39:5826–5836. [PubMed: 21459850]

72. Livi CM, Blanzieri E. Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. BMC Bioinformatics. 2014; 15:123. [PubMed: 24780077]

73. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein-RNA interactions. J Struct Biol. 2012; 179:261–268. [PubMed: 22019768]

74. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006; 34:W369–W373. [PubMed: 16845028]

75. Tran NT, Huang CH. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biol Direct. 2014; 9:4. [PubMed: 24555784]

76. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–1190. [PubMed: 15173120]

77. Reyes-Herrera PH, Ficarra E. Computational Methods for CLIP-seq Data Processing. Bioinform Biol Insights. 2014; 8:199–207. [PubMed: 25336930]

78. Slattery M, et al. Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci. 2014; 39:381–399. [PubMed: 25129887]

79. Zhao Y, Stormo G. Jury remains out on simple model of transcription factors. Nature Biotechnol. 2011; 6:480–483. [PubMed: 21654662]

80. Zhao Y, Ruan S, Pandey M, Stormo GD. Improved models for transcription factor binding site identification using nonindependent interactions. Genetics. 2012; 191:781–790. [PubMed: 22505627]

81. Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One. 2010; 5:e9722. [PubMed: 20339533]

82. Bulyk MLJ, PL, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. 2002; 30:1255–1261. [PubMed: 11861919]

83. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. PLoS Comput Biol. 2013; 9:e1003214. [PubMed: 24039567]

84. Grau J, Posch S, Grosse I, Keilwagen J. A general approach for discriminative de novo motif discovery from high-throughput data. Nucleic Acids Res. 2013; 41:e197. [PubMed: 24057214]

85. Zhou Q, Liu JS. Extracting sequence features to predict protein-DNA interactions: a comparative study. Nucleic Acids Res. 2008; 36:4137–4148. [PubMed: 18556756]

86. Hooghe B, Broos S, van Roy F, De Bleser P. A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. Nucleic Acids Res. 2012; 40:e106. [PubMed: 22492513]

87. Ben-Gal I, et al. Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics. 2005; 21:2657–2666. [PubMed: 15797905]

88. Liu LA, Bradley P. Atomistic modeling of protein-DNA interaction specificity: progress and applications. Curr Opin Struct Biol. 2012; 22:397–405. [PubMed: 22796087]

89. Han A, et al. De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. PLoS Comput Biol. 2014; 10:e1003442. [PubMed: 24499931]

90. Buenrostro JD, et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nature Biotechnol. 2014; 32:562–568. This work introduces the RNA Array technology and measures binding and dissociation kinetics for large numbers of RNA sequence variants. [PubMed: 24727714]

91. SantaLucia JJ, Turner DH. Measuring the thermodynamics of RNA secondary structure formation. Biopolymers. 1997; 44:309–319. [PubMed: 9591481]

92. Forsdyke DR. Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues. J Theor Biol. 2007; 248:745–753. [PubMed: 17698086]

93. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. Nucleic Acids Res. 2012; 40:e54. [PubMed: 22241775]

94. Maenner S, Müller M, Fröhlich J, Langer D, Becker PB. ATP-dependent roX RNA remodeling by the helicase maleless enables specific association of MSL proteins. Mol Cell. 2013; 51:174–184. [PubMed: 23870143]

95. Smith JK, Hsieh J, Fierke CA. Importance of RNA-protein interactions in bacterial ribonuclease P structure and catalysis. Biopolymers. 2007; 87:329–338. [PubMed: 17868095]

96. Rueda D, Hsieh J, Day-Storms JJ, Fierke CA, Walter NG. The 5′ leader of precursor tRNAAsp bound to the Bacillus subtilis RNase P holoenzyme has an extended conformation. Biochemistry. 2005; 44:16130–16139. [PubMed: 16331973]

97. Snoussi K, Leroy JL. Imino proton exchange and base-pair kinetics in RNA duplexes. Biochemistry. 2001; 40:8898–8904. [PubMed: 11467951]

98. Faustino I, Pérez A, Orozco M. Toward a consensus view of duplex RNA flexibility. Biophys J. 2010; 99:1876–1885. [PubMed: 20858433]

99. Bothe JR, et al. Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. Nature Methods. 2011; 8:919–931. [PubMed: 22036746]

100. Masliah G, Barraud P, Allain FH. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cell Mol Life Sci. 2013; 70:1875–1895. [PubMed: 22918483]

101. Cornish-Bowden A. Enzyme specificity: it's meaning in the general case. J Theor Biol. 1984; 108:451–457. [PubMed: 6748701]

102. Li J, et al. Identifying mRNA sequence elements for target recognition by human Argonaute proteins. Genome Res. 2014; 24:775–785. [PubMed: 24663241]

103. Lambert N, et al. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol Cell. 2014; 54:887–900. This paper introduces the RNA Bind-n-Seq method. [PubMed: 24837674]

104. Zearfoss NR, et al. A Conserved Three-Nucleotide Core Motif Defines Musashi RNA-Binding Specificity. J Biol Chem. 2014 epub.

105. Herschlag D. Implications of ribozyme kinetics for targeting the cleavage of specific RNA molecules in vivo: more isn't always better. Proc Natl Acad Sci U S A. 1991; 88:6921–9625. [PubMed: 1871108]

106. Auweter SD, Oberstrass FC, Allain FH. Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res. 2006; 34:4943–4959. [PubMed: 16982642]

107. Lamichhane R, et al. RNA looping by PTB: Evidence using FRET and NMR spectroscopy for a role in splicing repression. Proc Natl Acad Sci U S A. 2010; 107:4105–4110. [PubMed: 20160105]

108. Mickleburgh I, et al. The organization of RNA contacts by PTB for regulation of FAS splicing. Nucleic Acids Res. 2014; 42:8605–8620. [PubMed: 24957602]

109. Zhang W, et al. Crystal structures and RNA-binding properties of the RNA recognition motifs of heterogeneous nuclear ribonucleoprotein L: insights into its roles in alternative splicing regulation. J Biol Chem. 2013; 288:22636–22649. [PubMed: 23782695]

110. Romanelli MG, Diani E, Lievens PM. New insights into functional roles of the polypyrimidine tract-binding protein. Int J Mol Sci. 2013; 14:22906–22932. [PubMed: 24264039]

111. LaRiviere FJ, Wolfson AD, Uhlenbeck OC. Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. Science. 2001; 294:165–168. This paper introduces the concept of thermodynamic compensation for RBDs to achieve the nearly uniform affinity of eF-Tu for diverse RNAs. [PubMed: 11588263]

112. Nilsson J, Nissen P. Elongation factors on the ribosome. Curr Opin Struct Biol. 2005; 15:349–354. [PubMed: 15922593]

113. Hennig J, et al. Structural basis for the assembly of the Sxl-Unr translation regulatory complex. Nature. 2014; 515:287–290. [PubMed: 25209665]

114. Wasmuth EV, Januszyk K, Lima CD. Structure of an Rrp6-RNA exosome complex bound to poly(A) RNA. Nature. 2014; 511:435–439. [PubMed: 25043052]

115. Andersen CB, et al. Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. Science. 2006; 313:1968–1972. [PubMed: 16931718]

116. Wahl MC, Will CL, Lührmann R. The spliceosome: design principles of a dynamic RNP machine. Cell. 2009; 136:701–718. [PubMed: 19239890]

117. Zhou L, et al. Crystal structures of the Lsm complex bound to the 3′ end sequence of U6 small nuclear RNA. Nature. 2014; 506:116–120. [PubMed: 24240276]

118. Weber G, Trowitzsch S, Kastner B, Lührmann R, Wahl MC. Functional organization of the Sm core in the crystal structure of human U1 snRNP. EMBO J. 2010; 29:4172–4184. [PubMed: 21113136]

119. Kondo Y, Oubridge C, van Roon AM, Nagai K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. Elife. 2015:4.

120. Montemayor EJ, et al. Core structure of the U6 small nuclear ribonucleoprotein at 1.7-Å resolution. Nat Struct Mol Biol. 2014; 21:544–551. [PubMed: 24837192]

121. Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009; 136:731–745. [PubMed: 19239892]

122. Erzberger JP, et al. Molecular architecture of the 40S·eIF1·eIF3 translation initiation complex. Cell. 2014; 158:1123–1135. [PubMed: 25171412]

123. Marintchev A, et al. Topology and regulation of the human eIF4A/4G/4H helicase complex in translation initiation. Cell. 2009; 136:447–460. [PubMed: 19203580]

124. Antson AA, et al. Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. Nature. 1999; 401:235–242. [PubMed: 10499579]

125. Cieniková Z, Damberger FF, Hall J, Allain FH, Maris C. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. J Am Chem Soc. 2014; 136:14536–14544. This study correlates structures of multiple substrates with specificity information from high throughput studies of the RNA targets of the RBP in vivo. [PubMed: 25216038]

126. Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature. 2014; 505:701–705. [PubMed: 24336214]

127. Ding Y, et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014; 505:696–700. [PubMed: 24270811]

128. Wan Y, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature. 2014; 505:706–709. [PubMed: 24476892]

129. Ray D, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol. 2009; 27:667–670. [PubMed: 19561594]

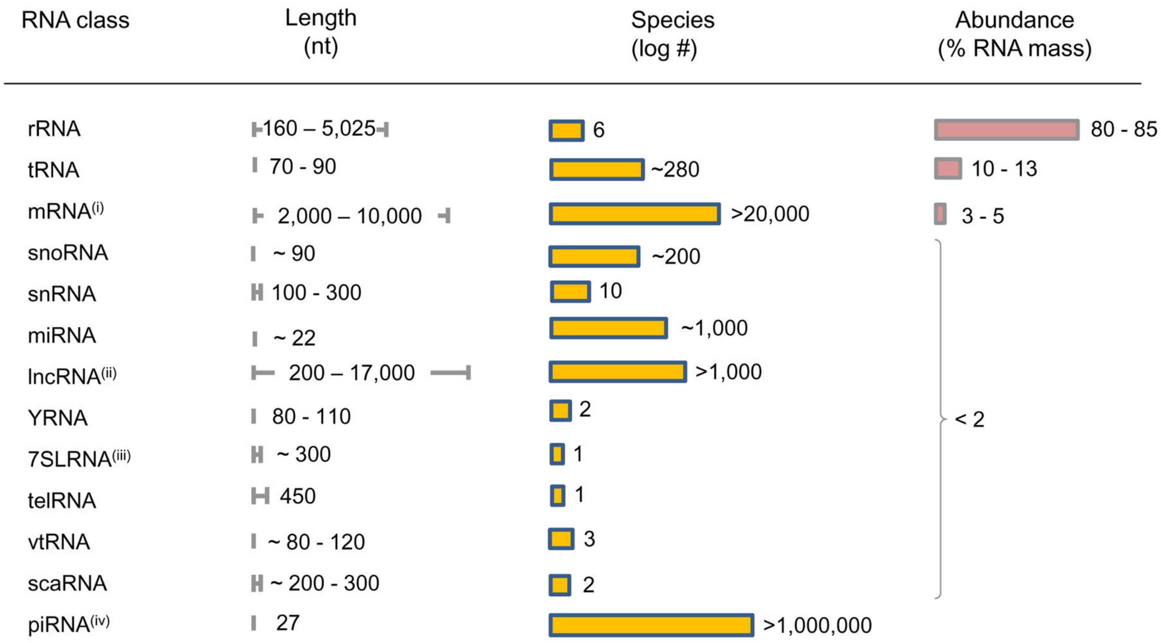
130. Ray D, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499:172–177. Refs 129 and 130 introduce and use the RNA Compete technology to define high affinity motifs for large numbers of RBDs. [PubMed: 23846655]

131. Tome JM, et al. Comprehensive analysis of RNA-protein interactions by high-thoughput sequencing-RNA affinity profiling. Nature Methods. 2014; 11:683–688. This paper introduces the HiTS-RAP technique. [PubMed: 24809628]

132. Roy R, Hohng S, Ha T. A practical guide to single-molecule FRET. Nature Methods. 2008; 5:507–516. [PubMed: 18511918]

133. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014; 157:77–94. [PubMed: 24679528]

134. Motorin Y, Helm M. RNA nucleotide methylation. Wiley Interdiscip Rev RNA. 2011; 2:611–631. [PubMed: 21823225]



RNA class	Length (nt)	Species (log #)	Abundance (% RNA mass)
rRNA	160 – 5,025	6	80 - 85
tRNA	70 - 90	~280	10 - 13
mRNA(i)	2,000 – 10,000	>20,000	3 - 5
snoRNA	~ 90	~200	< 2
snRNA	100 - 300	10	
miRNA	~ 22	~1,000	
lncRNA(ii)	200 – 17,000	>1,000	
YRNA	80 - 110	2	
7SLRNA(iii)	~ 300	1	
telRNA	450	1	
vtRNA	~ 80 - 120	3	
scaRNA	~ 200 - 300	2	
piRNA(iv)	27	>1,000,000	

Figure 1. The major classes of eukaryotic RNAs

For each class of RNA, the approximate length, number of different species and abundance are indicated. For more detailed information see REF. [133]. (*i*) The length of mRNAs reflects mature, processed mRNAs; the number of mRNA species refers to putative mRNA coding genes. (*ii*) long non coding RNAs (lncRNA) include all RNAs that do not explicitly belong to another RNA class, and exceed 200 nt in length. (*iii*) 7SLRNA refers to the RNA component of the signal recognition particle (SRP). (*iv*) piRNAs are expressed only at specific stages of germ cell development, and are not included in calculations of cellular RNA abundances.

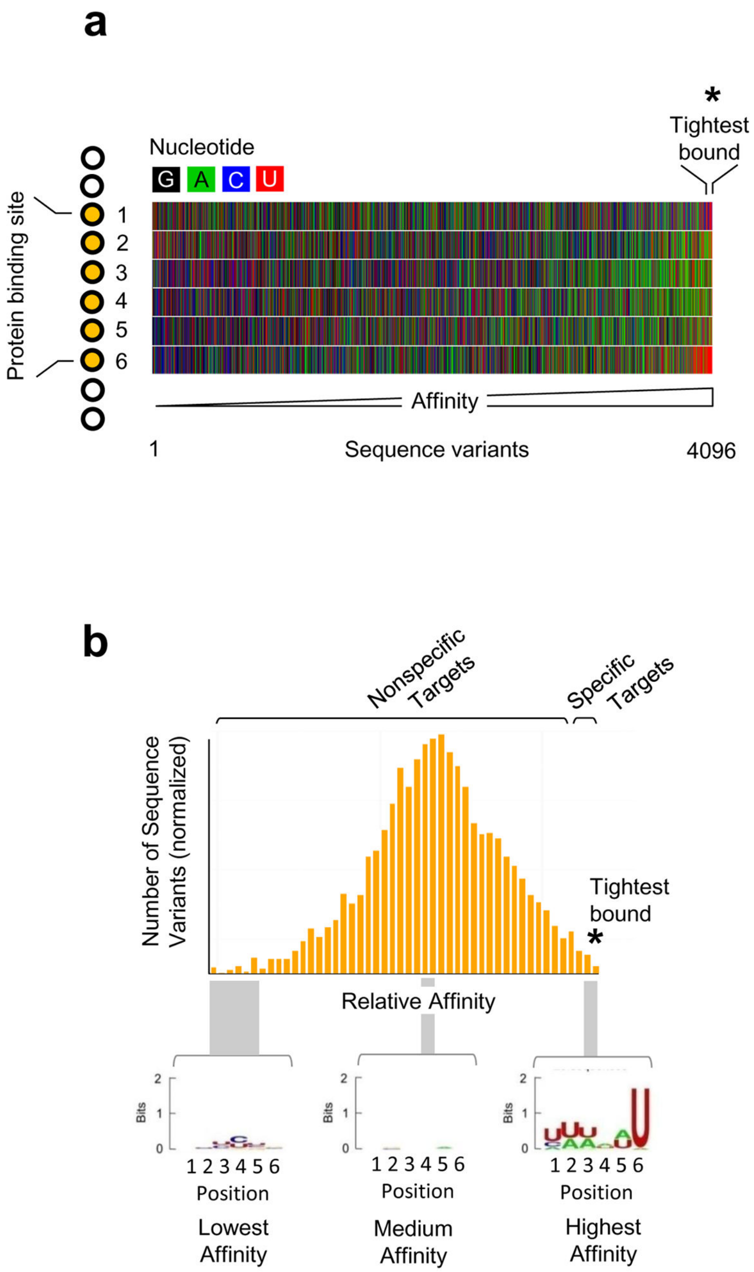


Figure 2. RBP affinity distributions

(**a**) Ranked affinities for an RBP with a binding site of 6 nucleotides (C5 from *E. coli*) to all RNA variants [67]. The numbers on the left indicate the nucleotide position in the binding site. (**b**) Histogram of relative affinities (log scale) for the sequence variants shown in panel (a). Relative affinities are calculated in relation to a standard variant, which can be chosen freely [67]. “Specific” RNA variants are marked by the asterisk and cluster in the high affinity region of the distribution and produce a binding consensus sequence (motif), shown as a logo underneath the plot. The remainder of the distribution consists of “non-specific” RNA variants, which do not produce a consensus motif.

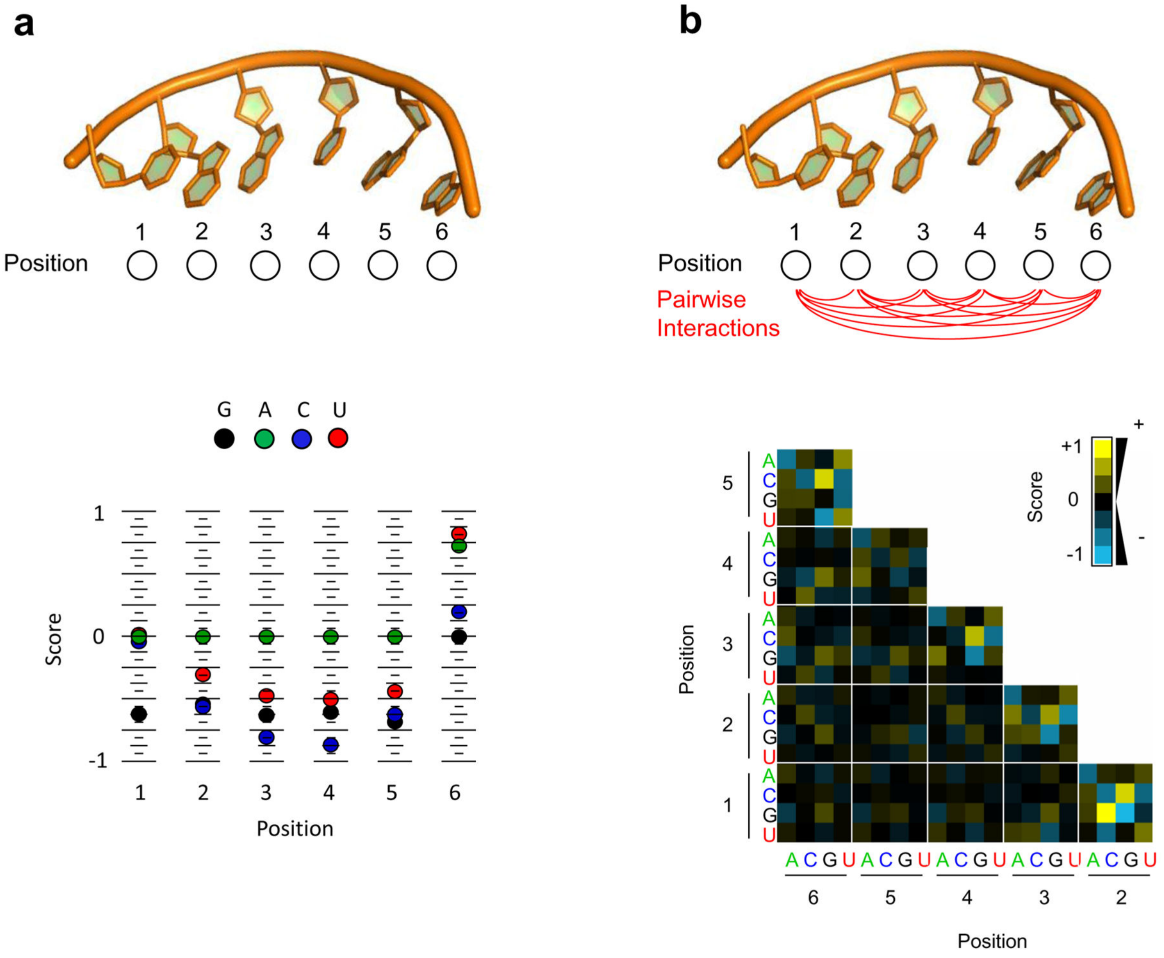


Figure 3. RBP binding models

(a) Position Weight Matrix (PWM). The structure denotes a hypothetical RNA binding site with six nucleotides. The plot (colored dots) depicts the score (linear coefficient) for each base at each position. The score is calculated from affinity distribution such as this shown in Figure 2(b). The score for each base corresponds to the contribution of the indicated nucleotide at each position to the overall binding free energy. (**b**) Binding model considering interactions between two bases (Pairwise Interaction Matrix - PIM, or Dinucleotide Weight Matrix - DWM). The structure denotes a hypothetical RNA binding site with six nucleotides, lines show the possible pairwise (energetic) couplings between two nucleotides. Colored fields correspond to the score for each of the 16 pairwise nucleotide permutation at each two positions. Scores are calculated from affinity distribution such as this shown in Figure 2(b). A yellow field (denoting a high score) indicates that a given nucleotide combination promotes binding (that is, increases the overall PWM score). A blue field (low score) indicates inhibition of binding by a given nucleotide combination. A black field indicates no significant contribution either way.

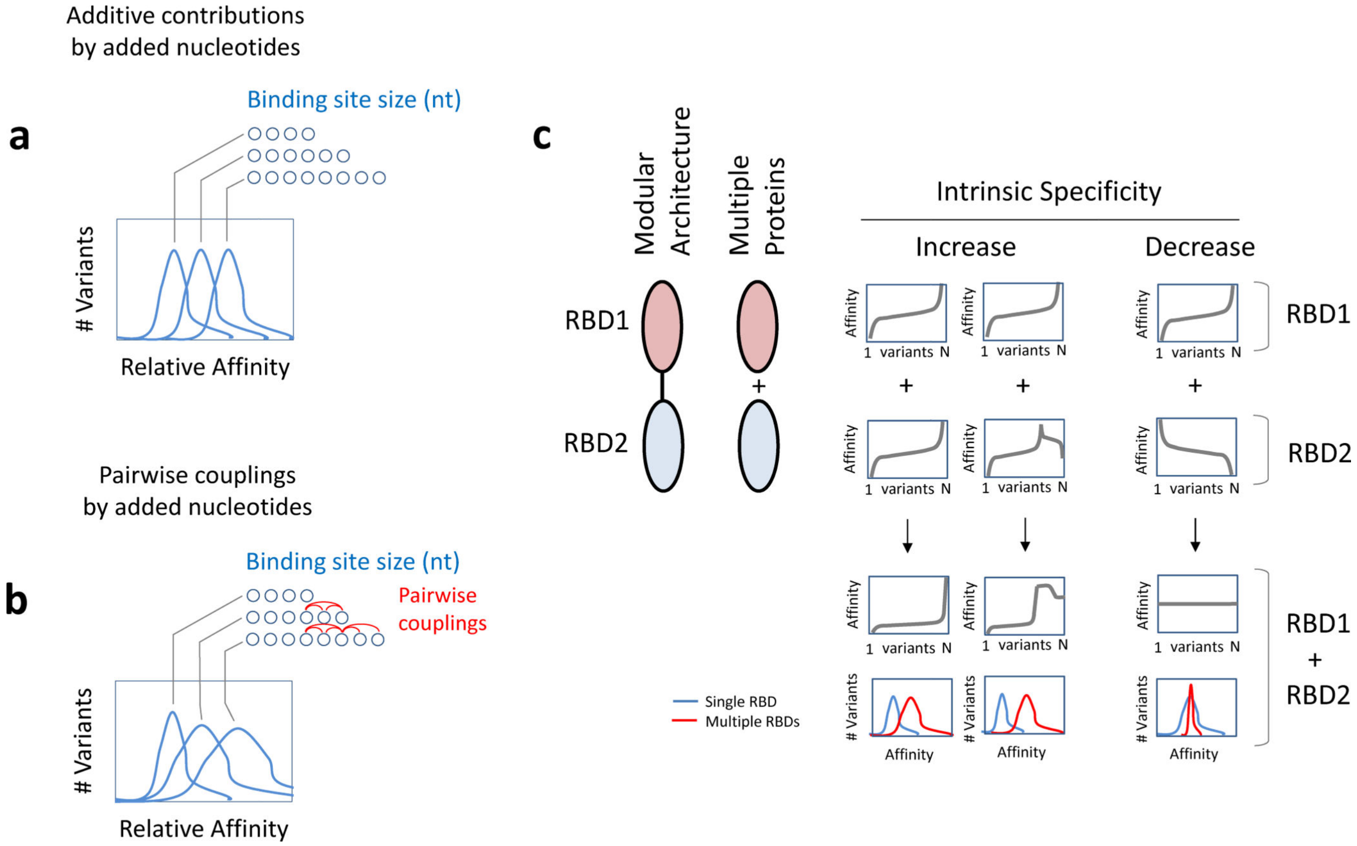


Figure 4. Strategies to increase or decrease intrinsic specificity of RBPs

(a) Change in binding site size with additive contributions by added nucleotides to binding energy. For a hypothetical RBP, additional nucleotides in a binding site would shift the affinity distribution towards higher affinities, but would not necessarily broaden the affinity distribution and thus not increase inherent specificity. (**b**) Change in binding site size with contributions of pairwise energetic couplings by added nucleotides. For a hypothetical RBP, hypothetical pairwise couplings by additional nucleotides in the binding site could strongly favor a small number of nucleotide combinations, thereby broaden the affinity distribution and thus greatly increase inherent specificity. **(c)** Increase or decrease in intrinsic specificity through multiple RBDs. Multiple RBDs (RBD1 and RBD2) can be part of the same protein or of separate proteins (left). The panels on the right show ranked affinity distributions (according to the same sequences for both RBDs) for each RDB. The panels in row three show the ranked affinity distribution upon combination of both RBDs, and the corresponding histogram of this ranked affinity distribution, color coded as indicated. Inherent protein specificity can be increased by additive specificities of the RBDs or decreased by compensatory specificities. Intrinsic specificities for individual RBDs can vary. Note, however, that binding preferences of individual RBDs do not need to be strictly additive, but can be synergistic, either through interactions between the RBDs or through cooperative binding of multiple several proteins.

Table 1
Classification of common protein domains that interact with RNA

The classification of ribonuclease domains is based on Anantharaman and Koonin[17], of helicase domains on Fairman *et al.*[25], and of methyltransferase domains on Motorin and Helm [134]. The compilation of the RNA-binding protein domains is based on Gerstberger *et. al.* [1].

Domain Class	Subclass (Superfamily)	Family
Nucleotidyl transferase	Poly(A) polymerases	Canonical PAPs
		Non-canonical PAPs
	Terminal UridylateTransferases	
	CCA-adding enzyme	
	Guanylyltransferase	
	RNA ligase	
	2′5′ poly(A) polymerase (OAS)	
	RNA-dependent RNA polymerase	
Ribonuclease	α/β	RNase A
		RNase H
		3 → 5 exo
		RNase II/R
		RNase E
		RNase PH
		Metallolactamase
	α + β	RNase T2
		XRN1
	α	RNase III
	Decapping enzyme	
RNA modifying enzymes	tRNAsynthetases	Class I
		Class II
	Deaminases	ADAR
		APOBEC
		TadA
		CDA
	Pseudouridine Synthases	
	Methyltransferases	RMFTase
		SPOUT
		Radical SAMTase
		FAD/NAD(p)
Helicase	SF1	Upf1-like
	SF2	Ski2-like
		RIG-I-like

Domain Class	Subclass (Superfamily)	Family
		DEAD-box
		DEAH/RHA
		Viral SF2
		Cas3
	SF3	
	SF4	
	SF5	
GTPase		EF-Tu/EF-G
		BMS1/SNU114
RNA-binding domains		RRM
		KH
		S1
		OB-fold
		PUF
		dsRBD
		Zn-fingers
		PAZ
		PIWI
		LSM
		KOW
		MIF4G
		NTF2
		GAR (RGG)
		HEAT repeat
		Homeodomain
		CSD