BioMol Concepts, Vol. 1 (2010), pp. 345–355 • Copyright © by Walter de Gruyter • Berlin • New York. DOI 10.1515/BMC.2010.037

Review

Dissection and prediction of RNA-binding sites on proteins

Laura Pérez-Cano and Juan Fernández-Recio*

Life Sciences Department, Barcelona Supercomputing Center (BSC), Jordi Girona 29, E-08034 Barcelona, Spain

*Corresponding author
e-mail: juanf@bsc.es

Abstract

RNA-binding proteins are involved in many important regulatory processes in cells and their study is essential for a complete understanding of living organisms. They show a large variability from both structural and functional points of view. However, several recent studies performed on protein-RNA crystal structures have revealed interesting common properties. RNA-binding sites usually constitute patches of positively charged or polar residues that make most of the specific and non-specific contacts with RNA. Negatively charged or aliphatic residues are less frequent at protein-RNA interfaces, although they can also be found either forming aliphatic and positive-negative pairs in protein RNA-binding sites or contacting RNA through their main chains. Aromatic residues found within these interfaces are usually involved in specific base recognition at RNA single-strand regions. This specific recognition, in combination with structural complementarity, represents the key source for specificity in protein-RNA association. From all this knowledge, a variety of computational methods for prediction of RNA-binding sites have been developed based either on protein sequence or on protein structure. Some reported methods are really successful in the identification of RNA-binding proteins or the prediction of RNA-binding sites. Given the growing interest in the field, all these studies and prediction methods will undoubtedly contribute to the identification and comprehension of protein-RNA interactions.

Keywords: computer prediction methods; interface propensities; protein-RNA interactions; RNA-binding protein; RNA-binding site.

Introduction

RNA-binding proteins (RBPs) play a crucial role in eukaryotic cells. They are involved in a diversity of processes within the gene expression pathways, by regulating the biogenesis, stability, transport, localization and function of RNAs (1–7). Furthermore, RBPs constitute a key target for medical and pharmaceutical purposes, especially considering that protein-RNA interactions are frequently involved in viral recognition and replication.

There are two main binding modes found in RBPs: (i) a secondary structure element binds into an RNA helix groove, and (ii) a β-sheet surface binds specifically to a single-stranded RNA (ssRNA) region (8). However, within each of these binding modes, there is a great diversity of interactions. It is supposed that the ability of RNAs to adopt alternative structures in combination to the relative resistance of nucleotides to mutation could have facilitated the evolution of new RNA-protein interactions (9). In fact, approximately 1500 human proteins are estimated to interact with RNA (5), considering that probably many RBPs have not yet been identified, based on the high number of proteins that remain without functional annotation.

In spite of this, since the first protein-RNA structure was characterized by Steitz and co-workers in 1989 (10), the growing awareness for the importance of RNA in the context of protein-RNA interactions, together with the publication of the 50S and 30S ribosome subunits in 2000 (11, 12), have increased the volume of structural data on this type of complexes. Some RNA-binding domains have already been well characterized (7, 13–22), although the number of high-resolution structures of protein-RNA complexes is still somewhat poor in comparison to that of other biomolecules. Some studies based on the available structural data of real protein-RNA interfaces have propelled the raising of general themes regarding the nature and mechanism of protein-RNA recognition. At the same time, all these data have also been used to develop different computational methods to predict new RBPs and RNA-binding sites and therefore to contribute to the protein-RNA identification and comprehension process.

In this review, we summarize the conclusions extracted from recent studies on RNA-binding sites and discuss the efficiency of many computational methods for the identification of RBPs and the prediction of RNA-binding sites on proteins, either using protein sequences or protein structures as prediction inputs.

Physicochemical properties of RNA-binding sites

Size of protein-RNA interfaces

In general, protein-RNA interfaces are larger than transient protein-protein interfaces, but smaller than permanent protein-protein interfaces or protein-DNA ones. However, owing to the variability of RNA secondary structure, a wide range of protein-RNA interface sizes can be found depending on the RNA type (23–27). Furthermore, even within the same RNA type the interface size can largely vary (24). For instance, protein-RNA complexes involved in protein syn-

thesis (basically tRNA/rRNA-protein complexes) have rather large interfaces (24, 25, 27), comparable to those in protein-DNA complexes as well as to homodimeric proteins (permanent assemblies) (24). By contrast, small protein-RNA interfaces, with similar size to those found in protein-protein transient interactions, can bind all types of RNA (25). Indeed, most of the protein-ssRNA complexes are included in this category (23). The large range of size values in protein-RNA interfaces is a reflection of the high structural variability that can be observed in protein-RNA complexes, which underlines the importance of the structural features in the RNA recognition process.

Amino acid composition at RNA-binding sites

The increasing number of available high-resolution structures of protein-RNA complexes has propelled the study of the evolutionary tendencies of protein residues to be involved in RNA-binding. From the different reported statistical analyses (Table 1) (23–32), it is possible to find general trends. All the reported studies found that the most preferred residues for RNA-binding are Arg, which is consistent with the high variety of RBPs containing the Arginine-rich RNA binding motif (9), as well as Lys. Other frequent residues in protein-RNA interfaces are His, Asn, Tyr and Ser. Regarding disfavored residues, most of the studies found the negatively charged residues Glu and Asp, as well as the aliphatic residues Ala, Val, Leu and Ile and Cys, as clearly not favored at RNA-binding sites. However, some clear inconsistencies were also observed as a result of the different size and composition of the data sets used to derive the propensities (23, 33). Aromatic residues gave the most inconsistent results. Among them, Phe was found as a preferred residue in three studies, but was disfavored in another three. Similarly, three studies found the Trp residue as favored, whereas another one found it as unfavorable. Interestingly, the study based on the largest data set reported different propensity values for aromatic residues when double-stranded RNA (dsRNA) and ssRNA complexes were considered separately, with aromatic residues significantly favored in ssRNA complexes (23). This was also reflected in another report, in which Phe and Tyr were favored from a data set where most of the protein-RNA complexes were of β-sheet type (β-sheet surface is known to specifically to bind a ssRNA region) (8). Thus, aromatic residues are expected to be more favored in complexes with mRNA, tRNA and vRNA, which have larger sections of ssRNA in comparison to those in rRNA (31). In addition, aromatic propensities are also more susceptible to depend on the definition of the propensity calculation. The propensity for a given residue is calculated as the observed frequency, computed from the relative residue composition of the protein-RNA interfaces, divided by the expected frequency, usually computed from the relative residue composition of protein surfaces. However, in the few studies in which Phe was found to be disfavored, the expected frequencies were computed based on the composition of the global protein sequences (29, 30, 32), which include buried residues and thus are more enriched in aromatic residues than the protein surfaces.

In general, these analyses indicate that aromatic residues seem to play an important role in the specific RNA recognition at single-stranded regions. A special scenario was found for the Tyr residue, which was mostly a preferred residue in all the different reported studies, even in those from databases enriched in dsRNA complexes. This could be explained by the polarity of its aromatic side chain, which allows Tyr to contact RNA bases not only through stacking interactions but also through both direct and water-mediated H-bonds (28).

Role of charged, polar, non-polar and aromatic residues

From the reported statistical propensities of protein residues in RNA-binding sites, one important conclusion emerges: positively charged and polar residues have a strong tendency to interact with RNA, whereas negatively charged and aliphatic residues are less likely to be involved in protein-RNA interfaces. In addition, aromatic residues are especially favored to contact ssRNA regions. Regarding the structural and physicochemical role of all the above described types of residues, positively charged residues Arg and Lys are the ones that most frequently make H-bonds to the O-1P/O-2P atoms in RNA molecules (34). Furthermore, the side chains of these two residues, together with all residue main chain NH groups, form half of the H-bonds contacts in protein-RNA interfaces (24). Arg and Lys also significantly participate in the recognition of the 2′-OH, although the group most commonly involved in this recognition is the carbonyl oxygen atom (C=O) of the polypeptide backbone. This wide range of participation of Arg and Lys in the formation of H-bonds is consistent with the fact that their H-bond forming atoms (N in Arg and Lys) are often highly solvent-exposed, and consequently they are more accessible to form hydrogen bonds (35). Arg can also form stacking interactions through its ionized guanidinium group over the physiological pH range (34). By contrast, in non-cognate interfaces (e.g., generated by crystal packing), these residues also have high interface propensities and preferentially form H-bonds with RNA phosphate groups (36). All these results highlight the importance of electrostatics and H-bonding, facilitated by positively charged residues, in mediating both specific and nonspecific protein-RNA interactions. Another residue commonly found at interfaces, His, can be positively charged depending on the environment. In this context, it was observed that most of His residues contact the RNA backbone, with more than half of them in contact with the phosphate (23). By contrast, the binding of His to U nucleotide is favored through water-mediated bonding with RNA (28).

Aromatic residues have been found in a wide variety of RBPs contacting the RNA bases through their side chains (8). Among them, Phe is the aromatic residue that is most frequently found forming stacking interactions (24). In fact, in around half of the cases in which a Phe is found at a RNA-binding site, it is involved in stacking interactions (37). These types of interactions are supposed to contribute to specificity, as was seen for Tyr and Phe, especially in ribonucleic recognition motifs (38). This is in agreement with

Table 1 Reported residue propensities at protein-RNA interfaces.

	Jones et al. (27)		Treger and Westhof (26)	Jeong et al. (28)		Kim et al. (25)	Lejeune et al. (29)	Terribilini et al. (30)	Ellis et al. (31)	Bahadur et al. (24)	Cheng et al. (32)	Perez-Cano and Fernandez-Recio (23)
Data set size[a]	20		26	51		40	49	62	53	81	62	170
Preferred residues	Lys, Tyr, Phe, Ile, Arg		Arg, Asn, Ser, Lys	Arg[b] Lys Asn Ser	Thr[c] Ser Arg Tyr	Arg, Lys, Tyr, Met, His, Gly, Phe	Arg, Lys, Asn, His, Gln, Asp, Tyr	Arg, Lys, His, Trp, Tyr	Trp, Arg, His, Ser, Gly, Lys	Arg, Lys, Phe, Tyr, Trp	Arg, Asn, Gln, Gly, His and Lys	Arg, Lys, His, Asn, Ser
Disfavored residues	Cys, His, Asp, Pro, Ala[d]		Ala, Ile, Leu Val	Val Ile Ala	Cys Pro Ala	Glu, Asp, Cys, Thr, Ala, Gln, Asn[d]	Ala, Val, Ile, Leu, Cys, Met, Trp, Glu, Phe	Phe, Glu, Asp, Leu, Ile, Val, Ala	Val, Leu, Asp Glu	Asp, Glu, Ile, Leu, Met, Val	Ala, Asp, Glu, Ile, Leu, Phe, Val	Asp, Glu, Cys, Val, Leu, Ile
Residue-base preferences[e]	Arg-U[f] Arg-X Asn-G Asn-U Glu-G Gly-G Thr-A Tyr-X	Asn-U[g] Ile-G Phe-A Phe-G Phe-U Tyr-X Trp-G Met-G Cys-G	Arg-X Lys-X Met-X Phe-X Tyr-X Ile/Pro/ Ser-A Leu-C Asp-G Gly-G Asn-U	Arg-U Asn-U Thr-A Lys-A	Thr-A Tyr-C Asn-U	No data	Arg-C Arg-G Lys-C Arg-U Lys-A	No data	Lys-X Tyr-U Arg-X Phe-A Trp-G	No data	No data	Arg-X Lys-X His-X Asn-U Asn-A

[a]Number of PDB entries in the non-redundant data sets used to extract the propensity values.
[b]Preferences considering direct H-bonds interactions.
[c]Preferences considering water-mediated H-bonds.
[d]Disfavored residues are not directly cited by their authors, thus we selected those with the lowest propensity values.
[e]Residue preferences for the ribonucleotide backbone (i.e., the sugar or the phosphate) are indicated with an X, which means no RNA base discrimination.
[f]Preferences considering H-bonds.
[g]Preferences considering van der Waals contacts.

the finding that these two residues have a smaller propensity in nonspecific interfaces (e.g., generated by crystal packing) (36).

Regarding the least favored residues, Ile and the negatively charged residues Asp and Glu are sometimes found to be forming aliphatic and positive-negative pairs in RNA-binding sites. Finally, it is also possible to find these and other disfavored residues in RNA-binding sites, contacting the RNA molecules through their main chain (26).

The specificity determinants of protein-RNA interaction

Thermodynamic studies have shown that binding specificity is generally a function of several factors, including base-specific hydrogen bonds, non-polar contacts and mutual accommodation of the protein and RNA-binding surfaces (8). In general, specificity is directly related to RNA base recognition (27, 34), although this recognition not only depends on the chemical affinities but also on the three-dimensional (3D) structural arrangement of RNA bases and interaction residues (33, 34, 39). Regarding the chemical affinities, from the interface composition point of view, specificity in recognition is mostly lying on the protein side. Although protein residues show different tendencies to bind RNA, none of the four ribonucleotides show a significant global preference for binding proteins (23, 24, 26). In fact, RNA recognition can be largely mediated by the interactions of the amide and carbonyl groups of the protein backbone with the edge of the RNA base (34), where the specificity is achieved because of the strong geometric constraints required for the H-bond interactions (39). This shows the importance of the RNA structural diversity for specificity (33). By contrast, RNA bases have to be accessible to protein residues to be recognized. For that reason, specific RNA recognition is largely found through ssRNA structures or ssRNA regions such as stem loops, bulges or kinks (34). Interestingly, these are the least conserved regions along evolution. Similarly, residues forming H-bonds with the RNA bases are less conserved than those involved in H-bonds with the RNA backbone (40).

Preferences in residue-RNA base interactions

As discussed above, there is no preference for protein binding among the four types of ribonucleotides (23, 24, 26). However, the characteristic shape and H-bonding pattern of RNA bases make some proteins residues to favor contacts with certain specific bases (37). Different statistical studies have reported pairwise protein-RNA propensities using available X-ray structures of protein-RNA complexes (Table 1). Regarding the most favored residue-nucleotide pairs, it seems difficult to extract general conclusions, as a variety of values are reported by the different studies. The Asn/U pair shows the most consistent preference values. By contrast, although some specific preferences were also reported for Arg/U (in three studies) and Lys/A pairs (in two studies), most of the studies found Arg and Lys residues not to prefer any specific nucleotide, basically because they mostly interact with the phosphate group (26, 27, 31).

Structural complementarity in protein-RNA complexes

Structural complementarity is one of the most important specificity determinants in protein-RNA recognition, indeed much more important than in protein-protein or protein-DNA interactions. For comparison, in protein-protein recognition there is variety of residue-residue interactions that can be used to achieve specificity, and protein-DNA interactions are basically dominated by solvent accessibility of nucleotides and amino acid side chain interactions (8). By contrast, RNA recognition frequently occurs in non-canonical and single-strand-like structures that use a much wider range of interaction geometries (37), with stronger orientational constraints on the relative placement of H-bond atom pairs (39). Furthermore, base-specificity can be achieved using only the polypeptide backbone, i.e., carbonyl and amide groups, provided they are found in a suitable 3D context (34). This corroborates the important contribution of shape complementarity to specificity (24). Interestingly, it has been recently shown that it is possible to recognize near-native protein-RNA docking poses based only on structural complementarity (33). By contrast, the possibility of protein and RNA interacting molecules to adapt to each other by structural rearrangements makes possible a large variety of interactions with high specificity. In this way, structural changes induced by protein-RNA interaction allow protein domains with common structural frameworks to recognize different RNA molecules with similar specificities, as in the coat proteins of different ssRNA phages (41–47) or in the RNA-binding domain of the Jembrane disease virus Tat protein (9).

Identification of RNA-binding proteins

In recent years, a high diversity of protein-RNA interactions has been described. However, a significant number of RBPs are estimated to still remain unknown. An important question emerges: is it possible to distinguish RNA-binding areas on the protein surface from areas that bind other proteins or DNA? With this purpose, different computational methods have been developed to try to identify RBPs from protein sequences or structures.

Protein-RNA interactions are dominated by electrostatics, which underlines a major difference with regard to protein-protein association, where desolvation and hydrophobic effect seem much more important (23). In this regard, residue composition at RNA-binding sites in proteins seems to be distinguishable from protein-binding sites as the former are composed of mostly polar and positively charged residues (Table 1). By contrast, RNA-binding sites appear to be smaller and less polar than DNA-binding sites (9, 24), but both share some common structural frameworks that make them more difficult to be distinguished based only on statistical parameters (8, 48).

Table 2 Comparison of reported methods to identify RNA-binding proteins.

Method	Description	Prediction input	Test set	PPV[a]	Sensitivity[b]	Availability
SVMProt (48)	Support vector machine (SVM) considering physicochemical residue properties	Protein sequence	447 RBPs 4881 non-RBPs	69%	98%	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi
PFplus (49, 50)	Prediction of electrostatic patches on protein surfaces	Protein structure	76 RBPs 246 non-RBPs	71%	80%	http://pfp.technion.ac.il/
PIRaNhA (51)	SVM with PSSM, interface propensities, predicted accessibility and hydrophobicity	Protein sequence	134 RBPs 134 non-RBPs	74%	80%	http://www.bioinformatics.sussex.ac.uk/PIRANHA/
OPRA (23)	Prediction of protein-RNA propensity-based patches on protein surfaces	Protein structure	30 RBPs 30 non-RBPs	77%	80%	juanf@bsc.es

[a]Positive predictive value: percentage of correctly classified RBPs among total number of proteins predicted as RBPs.
[b]Percentage of correctly classified RBPs among the total number of RBPs within the test set.

Sequence-based versus structure-based computational methods

Different methods for identifying RBPs from a pull of RNA-binding and non-RBPs are summarized in Table 2 (23, 48, 49, 51). The predictive success of the methods is usually reported by means of negative specificity (i.e., the capability of identifying proteins that do not bind RNA) and sensitivity values. However, using the negative specificity as an evaluation parameter is not sufficiently informative owing to the fact that some of the reported test sets contain many more non-RNA binding proteins than RBPs. In these conditions, it is expectedly easier to predict non-binding proteins, and thus high negative specificity might be achieved just by chance. For that reason, we have recalculated here the positive predictive value for each method, based on the corresponding published data (Table 3). We believe that this statistical parameter better reflects the reliability of the predictions.

As mentioned above, RNA-binding interfaces show some general common characteristics, such as amino acid composition, charge, polarity and hydrophobicity (48). This wide range of properties could explain the success on the identification of RBPs that can be achieved by different methods (Figure 1). Similar results can be obtained by both sequence-based and structure-based methods (Table 2), with a very good average positive predictive value of 73% and a remarkable average sensitivity of 84.5%. Most of non-RNA binding proteins found in the test sets used by the different studies correspond to non-nucleic acid binding proteins or proteins that bind other proteins. This explains the efficiency of all these reported methods in distinguishing RBPs from protein-binding proteins. By contrast, it seems to be more difficult to distinguish RBPs from DNA-binding proteins. In one study, it was observed that many incorrectly predicted proteins were actually DNA-binding proteins (48), whereas a different study reported the incapability of its method to distinguish RBPs from DNA-binding proteins (49).

The methods entirely based on the protein sequence have the advantage of the applicability, because there are many more annotated sequences than structures, and show reasonable results when using multiple sequence alignments by the position specific scoring matrix (PSSM) (51). By contrast, the structure-based methods are not based on conservation or evolutionary linked proteins, which make them potentially more useful for the identification of RBPs with novel binding motifs.

Prediction of RNA-binding sites on proteins

The prediction of potential RNA-binding residues on proteins can contribute to characterize the structure and mechanism of protein-RNA interactions (52) and has practical applications for wet-lab experiments, which are currently performed by costly mutagenesis approaches. With this purpose, a variety of computational methods have been recently developed to identify RNA-binding sites from protein sequences or unbound structures.

Table 3 Performance evaluation parameters of methods for identification of RNA-binding proteins.

	TP+FN	TN+FP	TP	TN	FP	FN	PPV	Sn	Sp-
SVMProt	447	4881	437	4685	196	10	69[a]	98	96
Pfplus	76	246	60	222	24	16	71[a]	79	90
PirAnHA	134	134	107[a]	97[a]	37[a]	27[a]	74[a]	80	72

We have used the published data according to: TP, number of correct RNA-binding protein predictions; TN, number of correct non-RNA-binding protein predictions; FP, number of incorrect RNA-binding protein predictions; FN, number of incorrect non-RNA-binding protein predictions; TP+FN, total number of RNA-binding proteins; TN+FP, total number of non-RNA-binding proteins; Sp=(TN/TN+FP)×100; Sn=(TP/TP+FN)×100; PPV=(TP/TP+FP)×100.

[a]Values were not provided by their authors and were calculated by us.

Sequence-based versus structure-based computational methods

The predictive success rates of different methods for identifying RNA-binding residues are summarized in Table 4 (23, 25, 30, 32, 51, 53–62). The efficiency of the methods is usually reported as negative specificity and sensitivity values. The test sets used to derive the parameters in these methods contained many more negative residues (i.e., non-RNA binding) than positive residues (i.e., RNA-binding), thus negative specificities are not very meaningful (it is relatively easy to predict non-RNA binding residues in a test set enriched in this type of residues). Furthermore, for experimental purposes, it is usually more important to know the reliability of the predicted residues as accurately as possible. For that reason, we have recalculated here the positive predictive value for each method, based on the corresponding published data (Table 5). The methods compiled here have been reported to predict RNA-binding residues based on statistical properties from protein sequences or structures (Table 4). Most of these methods rely on the fact that RNA-binding is a cooperative phenomenon (25) and interactions need to be considered in a patch (23, 27, 34) where the residue composition (Table 1) as well as the type of environment is essential (30). Therefore, sequence-based methods usually explore groups or windows of sequence neighbor residues and, similarly, structure-based methods analyze groups or patches of geometrically neighboring residues. The methods based entirely on protein sequences have poorer results in the prediction of RNA-binding sites when they do not use PSSM-related or sequence conservation. Only one of them, BindN (53), is able to achieve a sensitivity value above 50%, but with rather poor positive predictive value (PPV). However, most of the sequence-based methods that also include PSSMs or sequence conservation as input parameter, such as RNAProB (32) and Pprint (55), achieve much better results. Interestingly, there is only one method based only on protein structure, OPRA (23), which is able to identify RNA-binding residues with very good PPV (see some examples in Figure 2).

In summary, sequence-based methods have the theoretical advantage of wider applicability, but in practice only those methods that use multiple sequence alignments by PSSMs are actually reporting good results. By contrast, the structure-based methods, although needing the availability of high resolution protein structures, they do not need to use sequence conservation or evolutionary linked proteins and, thereby, they can be potentially more useful in identifying RNA-binding sites in novel RBPs.

Expert opinion

The phenomenon of protein-RNA association is receiving increasing attention in structural biology. In spite of the high variability of RBP types, RNA-binding interfaces show general properties different from those of protein-protein interfaces. RNA-binding sites are preferentially composed of

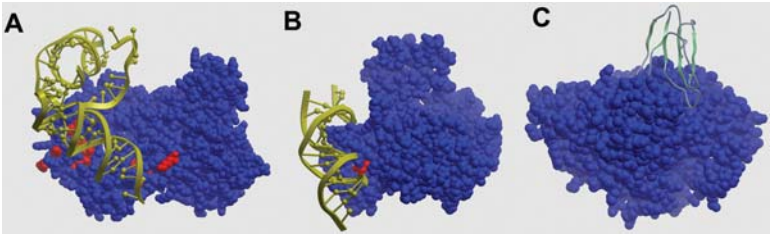

Figure 1 Identification of RNA-binding proteins by computer predictions.
Protein residues predicted by the OPRA predictor on the unbound protein are colored in red, indicating an RNA-binding signal. For comparison purposes, the position of the interacting molecule in the reference complex structure is also shown. (A) The RNA-binding residues predicted in the unbound archaeosine tRNA-guanine transglycolase (PDB 1IQ8) are correctly located in the RNA-binding interface (PDB 1J2B). (B) A small RNA-binding signal identified in the unbound catabolite gene activator protein (PDB 1I5Z) is found in the DNA-binding interface (PDB 1BER), which shows the difficulty to discriminate DNA-binding proteins from RBPs. (C) There is no RNA-binding signal in the unbound acetylcholinesterase (PDB 1J06), which can bind other proteins (PDB 1MAH) but it is not known to bind nucleic acids. This exemplifies the possibility of using the OPRA predictor to distinguish between RBPs and protein-binding proteins.

Table 4 Comparison of reported methods to predict RNA-binding sites on proteins.

Method	Description	Prediction input	Test set	PPV[a]	Sensitivity[b]	Availability
KYG (25)	Single and doublet propensities+PSSM applied to surface exposed residues	Protein sequence and structure	86 sequences	80%	10%	http://yayoi.kansai.jaea.go.jp/qbg/kyg
RNABindR (30)	Naive Bayes classifier	Protein sequence	109 sequences	51%	38%	http://bindr2.gdcb.iastate.edu/RNABindR/
BindN (53)	SVM considering three biochemical sequence-derived features	Protein sequence	107 sequences	28%	66%	http://bioinfo.ggc.org/bindn/
Chen and Lim's method (54)	Prediction of patches and/or clefts composed of residues with good electrostatics, conservation and solvent-accessible surface area	Protein sequence and structure	69 bound protein structures and their corresponding sequences	51%	37%	carmay@gate.sinica.edu.tw
RNAProB (32)	Smoothed PSSM considering the correlation and dependency from the neighboring residues	Protein sequence	86, 109 and 107 sequences	70% 64% 35%	80% 65% 77%	tsung@iis.sinica.edu.tw hsu@iis.sinica.edu.tw
Pprint (55)	SVM+PSSM profile	Protein sequence	86 and 107 sequences	61% 28%	53% 70%	http://www.imtech.res.in/raghava/pprint/
PRINTR (56)	SVM+PSSM profile+predicted secondary structure	Protein sequence	109 sequences	56%	46%	yanw@mail.hust.edu.cn
RISP (57)	SVM+PSSM profile	Protein sequence	147 and 71 sequences	36% 31%	61% 71%	zhlu@seu.edu.cn
PIRaNhA (51)	SVM with PSSM, interface propensities, predicted accessibility and hydrophobicity	Protein sequence	81 and 42 sequences	59% 48%	56% 53%	http://www.bioinformatics.sussex.ac.uk/PIRANHA/
PRIP (58)	SVM with PSSM, accessible surface area, betweenness-centrality and retention coefficient	Protein sequence and structure	144 bound protein structures and their corresponding sequences	28%	82%	z.yuan@imb.uq.edu.au
Struct-NB (59)	A classifier with sequence and structural features	Protein sequence and structure	147 bound protein structures and their corresponding sequences	44%	36%	ftowfic@cs.iastate.edu
RBRP (60)	Combining propensities, PSSM, secondary structure and solvent accessibility	Protein sequence	107 bound protein structures and their corresponding sequences	58%	52%	http://jeele.go.3322.org/RNA
BindN+ (61)	SVM with new descriptors of evolutionary information combined with PSSM	Protein sequence	107 sequences	37%	72%	http://bioinfo.ggc.org/bindn+/
OPRA (23)	Predictor of protein-RNA propensity-based patches on protein surfaces	Protein structure	30 unbound protein structures	74%	18%	juanf@bsc.es

[a]Positive predictive value: percentage of correctly classified RBPs among total number of proteins predicted as RBPs.
[b]Number of correctly classified RBPs among the total number of RBPs within the test set.

Table 5 Performance evaluation parameters of methods for prediction of RNA-binding residues.

	TP+FN	TN+FP	TP[a]	TN[a]	FP[a]	FN[a]	PPV[a]	Sn	Sp-
BindN	3239	18 519	2147	12 934	5585	1092	28	66	70
RNAProB [86]	4568	15 503	3652	14 009	1494	916	71	80	90
RNAProB [109]	3581	21 526	2314	20 209	1317	1267	64	65	94
RNAProB [107]	2555	19 496	1971	15 766	3730	584	35	77	81
Pprint [86]	4568	15 503	2423	13 883	1620	2145	60	53	90
Pprint [107]	2555	19 496	1693	13 616	5880	862	22	66	70
RISP [147]	4336	27 988	2645	23 314	4674	1691	36	61	83
RISP [71]	1810	13 668	1291	10 839	2829	519	31	71	79
PiRaNhA [81]	2938	16 175	1654	15 010	1165	1284	59	56	93
PiRaNhA [42]	1279	7298	678	6568	730	601	48	53	90
PRIP	4304	27 932	3529	18 659	9273	775	28	82	67
RBRP	3239	18 519	1697	17 297	1222	1542	58	52	93
BindN+	3239	18 519	2319	14 574	3945	920	37	72	79

We have used the published data according to: TP, number of correct RNA-binding residue predictions; TN, number of correct non-RNA-binding residue predictions; FP, number of incorrect RNA-binding residue predictions; FN, number of incorrect non-RNA-binding residue predictions; TP+FN, total number of RNA-binding residues; TN+FP, total number of non-RNA-binding residues; Sp=(TN/TN+FP)×100; Sn=(TP/TP+FN)×100; PPV=(TP/TP+FP)×100.

[a]Values were not provided by their authors and were calculated by us.

positively charged and polar residues such as Arg, Lys, Ser and His. Thus, electrostatics and H-bonding mediated by positively charged residues are important in both specific and nonspecific protein-RNA interactions. However, there is no clear consensus about the preferences of aromatic residues for protein-RNA interfaces. Contradictory statistical conclusions can be extracted from data sets containing different proportions of single-stranded RNA structures. This emphasizes the importance of choosing a suitable data set to derive statistical propensities and warns about the interpretation of reported propensity values. Of course, it would be desirable to establish common standards for statistical studies, but in practical terms the conclusion from the currently available data is that the structural, functional and mechanistic characteristics of RBPs are going to strongly depend on the type of RNA they bind.

All these studies give insights about specificity of protein-RNA binding. It has been found that aromatic residues play an important role in specificity by making stacking interactions with unpaired bases. Indeed, specific recognition of RNA bases commonly occurs in single-stranded RNA regions, which interestingly are the least conserved regions along evolution. By contrast, specificity can also be achieved by protein-RNA structural complementarity, which frequently involves structural changes upon binding. Structural complementarity in protein-RNA binding is much more important than in protein-protein recognition, somehow expected from the large conformational flexibility of RNA molecules.

Based on the reported structural and physicochemical studies, several computational methods have been developed to identify RBPs and to predict RNA-binding sites on proteins. Some prediction methods are entirely based on sequence information, whereas others need some type of structural information for their predictions. Sequence-based methods are of more general applicability, although they need the inclusion of evolutionary information through PSSMs, whereas structure-based methods could be more suitable to identify novel RBPs and RNA-binding sites that are not yet annotated in functional databases. However, the comparison between predictive methods is currently difficult, because the reported results are based on different statistical evaluation measures. It would be important to achieve a consensus for the evaluation of predictive success rates and to establish common assessment experiments in the spirit of critical assessment of protein structure prediction (CASP) for structural prediction or critical assessment of prediction of interactions (CAPRI) for protein-protein prediction methods. New advances in RNA-binding prediction, in combination with structural data, are expected to boost the field of protein-RNA recognition.

Outlook

In recent years, the growing awareness for the importance of RNA and the publication of the 50S and 30S ribosome subunits (11, 12) have increased the amount of available data and, consequently, the knowledge about protein-RNA asso-

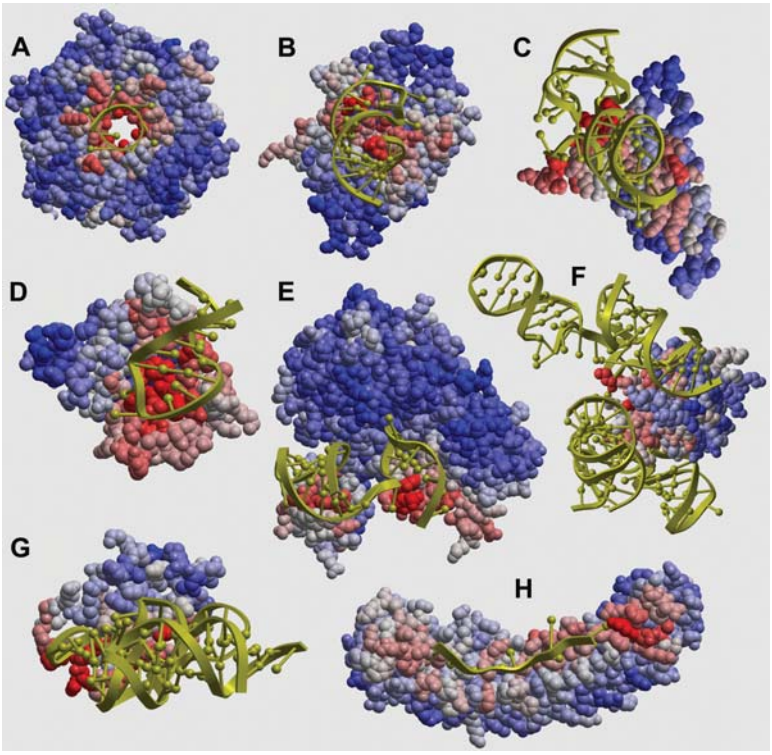

Figure 2 Examples of RNA-binding site predictions on protein surfaces.
Unbound protein residues predicted to be involved in RNA-binding according to the OPRA method are colored in red. For comparison purposes, the position of the RNA in the reference complex structure is shown in yellow ribbon. (A) Hfq pleiotropic translational regulator (unbound PDB 1KQ1; complex PDB 1KQ2). (B) PP7 coat dimer protein (unbound PDB 2QUD; complex PDB 2QUX). (C) SmpB tmRNA-binding protein from *Aquifex aeolicus* (unbound PDB 1K8H; complex PDB 1P6V). (D) *Aspergillus* ribotoxin (unbound PDB 1AQZ; complex PDB 1JBS). (E) 23S rRNA (Uracil-5-)Methyltransferase RUMA (unbound PDB 1UWV; complex PDB 2BH2). (F) SmpB tmRNA-binding protein from *Thermus thermophilus* (unbound PDB 1J1H; complex PDB 2CZJ). (G) L11 ribosomal binding domain (unbound PDB 1ACI; complex PDB 1HC8). (H) PUF4 protein (unbound PDB 3BWT; complex PDB 3BX2).

ciation. This has also propelled the development of computational methods that are able to identify novel RBPs and RNA-binding sites. Future methodological advances for computer characterization of RBPs and the introduction of molecular dynamics are expected to largely contribute to understand the functionality and mechanism of protein-RNA interactions and the important types of biomolecular processes in which they are involved.

Highlights

- There is a high variability of RBPs from structural and functional points of view, but their residue composition shows some common properties.
- RNA-binding sites are preferentially composed of positively charged and polar residues such as Arg, Lys, Ser, His and rarely composed of negatively charged and polar residues such as Glu, Asp, Ala, Val, Leu, Cys and Ile.
- Specificity is given by both base specific recognition and protein-RNA structural complementarity.
- Specific recognition of RNA bases usually occurs by interaction of aromatic residues to single-stranded RNA regions, which interestingly are the least conserved regions along evolution.
- The lack of clear consensus about the preference of some aromatic residues for protein-RNA interfaces is due to the different composition in ssRNA of the data sets used to derive the statistical analysis.
- Sequence-based and structure-based computational methods can significantly contribute to identifying RBPs as well as RNA-binding sites.
- Sequence-based methods need the use of evolutionary information to show successful predictions.
- Structure-based methods are suitable to identify novel RBPs and RNA-binding sites, although they need a high resolution protein structure.
- Identifying novel RBPs and RNA-binding sites will contribute to a better understanding of protein-RNA association, which in turn will improve current computational methods for better predictions.

References

1. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008; 582: 1977–86.

2. Lee MH, Schedl T. RNA-binding proteins. WormBook 2006: 1–13.
3. Chen Y, Varani G. Protein families and RNA recognition. FEBS J 2005; 272: 2088–97.
4. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 2002; 30: 1427–64.
5. Keene JD. Ribonucleoprotein infrastructure regulating the flow of genetic information between the genome and the proteome. Proc Natl Acad Sci USA 2001; 98: 7018–24.
6. Lasko P. The Drosophila melanogaster genome: translation factors and RNA binding proteins. J Cell Biol 2000; 150: F51–6.
7. Burd CG, Dreyfuss G. Conserved structures and diversity of functions of RNA-binding proteins. Science 1994; 265: 615–21.
8. Draper DE. Themes in RNA-protein recognition. J Mol Biol 1999; 293: 255–70.
9. Smith CA, Calabro V, Frankel AD. An RNA-binding chameleon. Mol Cell 2000; 6: 1067–76.
10. Rould MA, Perona JJ, Soll D, Steitz TA. Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 Å resolution. Science 1989; 246: 1135–42.
11. Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. Structure of the 30S ribosomal subunit. Nature 2000; 407: 327–39.
12. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 2000; 289: 905–20.
13. Frazao C, McVey CE, Amblar M, Barbas A, Vonrhein C, Arraiano CM, Carrondo MA. Unravelling the dynamics of RNA degradation by ribonuclease II and its RNA-bound complex. Nature 2006; 443: 110–4.
14. Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FH. Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. Nat Struct Mol Biol 2006; 13: 160–7.
15. Beuth B, Pennell S, Arnvig KB, Martin SR, Taylor IA. Structure of a *Mycobacterium tuberculosis* NusA-RNA complex. EMBO J 2005; 24: 3576–87.
16. Ma JB, Yuan YR, Meister G, Pei Y, Tuschl T, Patel DJ. Structural basis for 5′-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature 2005; 434: 666–70.
17. Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. Nat Struct Mol Biol 2004; 11: 257–64.
18. Wang X, McLachlan J, Zamore PD, Hall TM. Modular recognition of RNA by a human pumilio-homology domain. Cell 2002; 110: 501–12.
19. Ramos A, Grunert S, Adams J, Micklem DR, Proctor MR, Freund S, Bycroft M, St Johnston D, Varani G. RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J 2000; 19: 997–1009.
20. Lewis HA, Musunuru K, Jensen KB, Edo C, Chen H, Darnell RB, Burley SK. Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. Cell 2000; 100: 323–32.
21. Antson AA, Dodson EJ, Dodson G, Greaves RB, Chen X, Gollnick P. Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. Nature 1999; 401: 235–42.
22. Oubridge C, Ito N, Evans PR, Teo CH, Nagai K. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature 1994; 372: 432–8.
23. Perez-Cano L, Fernandez-Recio J. Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins 2010; 78: 25–35.
24. Bahadur RP, Zacharias M, Janin J. Dissecting protein-RNA recognition sites. Nucleic Acids Res 2008; 36: 2705–16.
25. Kim OT, Yura K, Go N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. Nucleic Acids Res 2006; 34: 6450–60.
26. Treger M, Westhof E. Statistical analysis of atomic contacts at RNA-protein interfaces. J Mol Recognit 2001; 14: 199–214.
27. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein-RNA interactions: a structural analysis. Nucleic Acids Res 2001; 29: 943–54.
28. Jeong E, Kim H, Lee SW, Han K. Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. Mol Cell 2003; 16: 161–7.
29. Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins 2005; 61: 258–71.
30. Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D. Prediction of RNA binding sites in proteins from amino acid sequence. RNA 2006; 12: 1450–62.
31. Ellis JJ, Broom M, Jones S. Protein-RNA interactions: structural analysis and functional classes. Proteins 2007; 66: 903–11.
32. Cheng CW, Su EC, Hwang JK, Sung TY, Hsu WL. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinformatics 2008; 9 (Suppl 12): S6.
33. Perez-Cano L, Solernou A, Pons C, Fernandez-Recio J. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Pac Symp Biocomput 2010: 293–301.
34. Allers J, Shamoo Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J Mol Biol 2001; 311: 75–86.
35. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994; 238: 777–93.
36. Phipps KR, Li H. Protein-RNA contacts at crystal packing surfaces. Proteins 2007; 67: 121–7.
37. Morozova N, Allers J, Myers J, Shamoo Y. Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. Bioinformatics 2006; 22: 2746–52.
38. Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 2005; 272: 2118–31.
39. Chen Y, Kortemme T, Robertson T, Baker D, Varani G. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. Nucleic Acids Res 2004; 32: 5147–62.
40. Spriggs RV, Jones S. RNA-binding residues in sequence space: conservation and interaction patterns. Comput Biol Chem 2009; 33: 397–403.
41. Spingola M, Lim F, Peabody DS. Recognition of diverse RNAs by a single protein structural framework. Arch Biochem Biophys 2002; 405: 122–9.
42. Tars K, Fridborg K, Bundule M, Liljas L. The three-dimensional structure of bacteriophage PP7 from *Pseudomonas aeruginosa* at 3.7-Å resolution. Virology 2000; 272: 331–7.
43. Tars K, Bundule M, Fridborg K, Liljas L. The crystal structure of bacteriophage GA and a comparison of bacteriophages

belonging to the major groups of *Escherichia coli* leviviruses. J Mol Biol 1997; 271: 759–73.
44. Golmohammadi R, Fridborg K, Bundule M, Valegard K, Liljas L. The crystal structure of bacteriophage Qβ at 3.5 Å resolution. Structure 1996; 4: 543–54.
45. Ni CZ, Syed R, Kodandapani R, Wickersham J, Peabody DS, Ely KR. Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly. Structure 1995; 3: 255–63.
46. Liljas L, Fridborg K, Valegard K, Bundule M, Pumpens P. Crystal structure of bacteriophage fr capsids at 3.5 Å resolution. J Mol Biol 1994; 244: 279–90.
47. Golmohammadi R, Valegard K, Fridborg K, Liljas L. The refined structure of bacteriophage MS2 at 2.8 Å resolution. J Mol Biol 1993; 234: 620–39.
48. Han LY, Cai CZ, Lo SL, Chung MC, Chen YZ. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 2004; 10: 355–68.
49. Shazman S, Mandel-Gutfreund Y. Classifying RNA-binding proteins based on electrostatic properties. PLoS Comput Biol 2008; 4: e1000146.
50. Shazman S, Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y. Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. Nucleic Acids Res 2007; 35: W526–30.
51. Spriggs RV, Murakami Y, Nakamura H, Jones S. Protein function annotation from sequence: prediction of residues interacting with RNA. Bioinformatics 2009; 25: 1492–7.
52. Eliahoo E, Ben Yosef R, Perez-Cano L, Fernandez-Recio J, Glaser F, Manor H. Mapping of interaction sites of the Schizosaccharomyces pombe protein Translin with nucleic acids and proteins: a combined molecular genetics and bioinformatics study. Nucleic Acids Res 2010; 38: 2975–89.
53. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 2006; 34: W243–8.
54. Chen YC, Lim C. Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res 2008; 36: e29.
55. Kumar M, Gromiha MM, Raghava GP. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 2008; 71: 189–94.
56. Wang Y, Xue Z, Shen G, Xu J. PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. Amino Acids 2008; 35: 295–302.
57. Tong J, Jiang P, Lu ZH. RISP: a web-based server for prediction of RNA-binding sites in proteins. Comput Methods Programs Biomed 2008; 90: 148–53.
58. Maetschke SR, Yuan Z. Exploiting structural and topological information to improve prediction of RNA-protein binding sites. BMC Bioinformatics 2009; 10: 341.
59. Towfic F, Caragea C, Gemperline DC, Dobbs D, Honavar V. Struct-NB: predicting protein-RNA binding sites using structural features. Int J Data Min Bioinform 2010; 4: 21–43.
60. Li Q, Cao Z, Liu H. Improve the prediction of RNA-binding residues using structural neighbours. Protein Pept Lett 2010; 17: 287–96.
61. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 2010; 4 (Suppl 1): S3.
62. Carson MB, Langlois R, Lu H. NAPS: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res 2010; 38 (Suppl): W431–5.