Advanced Review

# Predictions of protein–RNA interactions

Davide Cirillo, Federico Agostini and Gian Gaetano Tartaglia*

**Ribonucleoprotein interactions play important roles in a wide variety of cellular processes, ranging from transcriptional and posttranscriptional regulation of gene expression to host defense against pathogens. High throughput experiments to identify RNA–protein interactions provide information about the complexity of interaction networks, but require time and considerable efforts. Thus, there is need for reliable computational methods for predicting ribonucleoprotein interactions. In this review, we discuss a number of approaches that have been developed to predict the ability of proteins and RNA molecules to associate.**


**How to cite this article:**
*WIREs Comput Mol Sci* 2012. doi: 10.1002/wcms.1119

## INTRODUCTION

In the course of the past decade, a number of large consortia, most notably FANTOM[1] and ENCODE,[2] carried out large-scale sequencing of the human transcriptome. These and subsequent studies revealed a stunning and unexpected result—that the majority of the human genome is transcribed into many thousands of previously uncharacterized RNA molecules, both long and short. As a matter of fact, protein-coding genes occupy less than 2% of the human genome and represent dispersed oases in a landscape of DNA areas that are actively transcribed under developmental and environmental conditions. Most of the RNA molecules do not encode protein, and are transcribed from both 'intergenic' and previously described protein-coding loci. Noncoding RNAs (ncRNAs) can be classified by their size: 'short' RNAs, including well-known classes such as microRNAs or snoRNAs, and 'long' RNAs which are defined as any transcript >200 nucleotides that does not have a functional open reading frame. Conservative estimates have ~10,000 long ncRNAs (GENCODE),[3] and ~1500 microRNAs (mirBase Release 17). At present, the full repertoire of noncoding RNAs in the human genome is unclear—ever deeper RNAseq analyses of human transcriptomes does not appear close to saturation outside of protein-coding exons, suggesting that many more lowly expressed or cell-type-specific noncoding RNAs (ncRNAs) remain to be discovered, particularly in the poorly explored non-polyA fraction.[4] As for their function, it seems that most ncRNAs serve to regulate gene expression in some way by interacting with DNA and protein molecules.[5] In fact, it has been suggested that ncRNAs are slowly acquiring more *power* over evolution[6] and that ncRNA are taking on primary responsibility to regulate cellular processes and could co-opt proteins to assist in orchestrating the diversity of eukaryotic cellular pathways. Intriguingly, almost all known examples of ncRNA function by interfacing with protein complexes such as transcription, splicing, replication, and transport.[7,8] Indeed, the abundance and diversity of RNA-binding proteins (RBPs) has been correlated with the complexity of organisms, with the number of RBPs reaching thousands in vertebrates.[9]

There are numerous protein domains involved in RNA binding, with prevalence of $\alpha/\beta$ structures.[10] For instance, some common and well-characterized RNA-binding domains include: K Homology (KH) domain, Arginine Glycine Glycine (RGG) box, RNA Recognition Motif (RRM), double-stranded RNA-binding domain (dsRBD), Pumilio/FBF (PUF) domain, and the Piwi/Argonaute/Zwille (PAZ) domain.[7] The KH is the most abundant domain in structural databases (173 pdb entries) with a total of 16,184 KH regions identified in different proteomes.[11] By

*Correspondence to: gian.tartaglia@cgr.es
Centre for Genomic Regulation, CRG and UPF, Bioinformatics and Genomics, Dr. Aiguader 88, 08003 Barcelona, Spain
DOI: 10.1002/wcms.1119

  

contrast, only 20 pdb entries are present for the PUF domain.[11]

In general, computational predictions of RNA binding requires knowledge on whether a given protein binds RNA, which residues in the protein sequence are directly involved in making contacts with the RNA, which nucleotides interact with the protein and what is the structure of the protein–RNA complex. Availability of protein tertiary structure can greatly facilitate the prediction of RNA-binding sites, which is typically identified by surface-exposed residues that are close to each other in space, but not necessarily in sequence. RNA-binding sites are often positively charged patches exposed to the solvent to bind to negatively charged RNA backbone.[12] Structure-based predictive methods exploit the distribution of charged amino acids and spatial proximity of residues with particular features. Sequence-based algorithms employ the same information as structure-based methods, but replace structural observable with predicted features.

In this review, we discuss computational methods for the identification of protein and RNA-binding sites. Algorithms can be group in two classes depending on whether patterns derived from primary or tertiary structure are used for training. The catRAPID method is the first method that simultaneously predicts protein and RNA binding sites exploiting prediction of physicochemical properties such as secondary structure, hydrogen bonding and van der Waals propensities.

## PHYSICOCHEMICAL DETERMINANTS OF PROTEIN–RNA ASSOCIATIONS (catRAPID)

catRAPID (http://tartaglialab.crg.cat/) is the first computational method able to perform large-scale predictions of protein–RNA associations.[13] The algorithm was trained on a large set of protein–RNA pairs available in the Protein Data Bank to discriminate interacting and noninteracting molecules using secondary structure propensities, hydrogen bonding, and van der Waals contributions. Accurate predictions have been reported for long noncoding RNA associations with proteins and in particular the Polycomb chromatin-remodeling complex, suggesting that the approach could be particularly suitable for investigating RNA molecules involved in epigenetic regulation.[13] As interactions between proteins and long ncRNAs are known to play a pivotal role in gene transcription, histone and DNA methylation, acetylation, and sumoylation, catRAPID could be particularly suitable to predict lncRNA associations with RNA polymerases, transcription factors and chromatin modifiers to discover new protein–RNA interactions (Figure 1, Table 1).[14]

**TABLE 1** | The catRAPID Method

| |
|---|
| *Dataset*: 858 RNA–protein complexes available from the RCSB databank: a positive dataset containing 7409 interacting protein–RNA pairs and a negative set containing 958 noninteracting protein–RNA pairs. catRAPID is tested on the non-nucleic-acid-binding database (NNBP; area under the receiver operating characteristic (ROC) curve of 0.92), the NPInter database (area under the ROC curve of 0.88), and a number of individual interactions (e.g., RNAse mitochondrial RNA MRP and X-inactive specific transcript XIST networks). Performances are estimated with a 10-fold cross-validation. |
| *Method*: catRAPID is trained to discriminate between interacting and noninteracting molecules using secondary structure propensities, hydrogen bonding and van der Waals contributions. Physicochemical properties are combined into the interaction profiles, from which protein and RNA-binding sites are calculated. |

## Examples of Predictions

### *HOTAIR and SUZ12*

HOTAIR (HOX Antisense Intergenic RNA) is a 2.3 kb long intergenic ncRNA that serves as a scaffold for histone modification complexes.[15] The 5′ domain of HOTAIR binds Polycomb Repressive Complex 2 (PRC2), whereas the 3′ domain does not. HOTAIR is shuttled from chromosome 12 to chromosome 2 by the Suppressor of Zeste Homolog SUZ12 protein. By contrast, indoleglycerol phosphate synthase, which belongs to the class non-nucleic-acid binding proteins (NNBP)[12] shows negligible propensity to bind to the 5′ of HOTAIR RNA (discriminative power 0%).[11,12] Moreover, Suz12 does not bind to HOTAIR 3′ (discriminative power 0%), as previously reported.[15]

### *The RNP Complex*

There are many clear examples of noncoding RNAs that function as part of ribonucleoprotein (RNP) complexes (e.g., ribosome, splicesome, SRP, and RNase P). A classic example of a functional RNP complex is RNase P,[16] which is found in all kingdoms of life and is responsible for the generation of mature 5′ ends of tRNAs by cleaving the 5′ leader elements of precursor tRNAs (pre-tRNAs). In bacteria and eukaryotes, RNase P is composed of a core RNA and

  

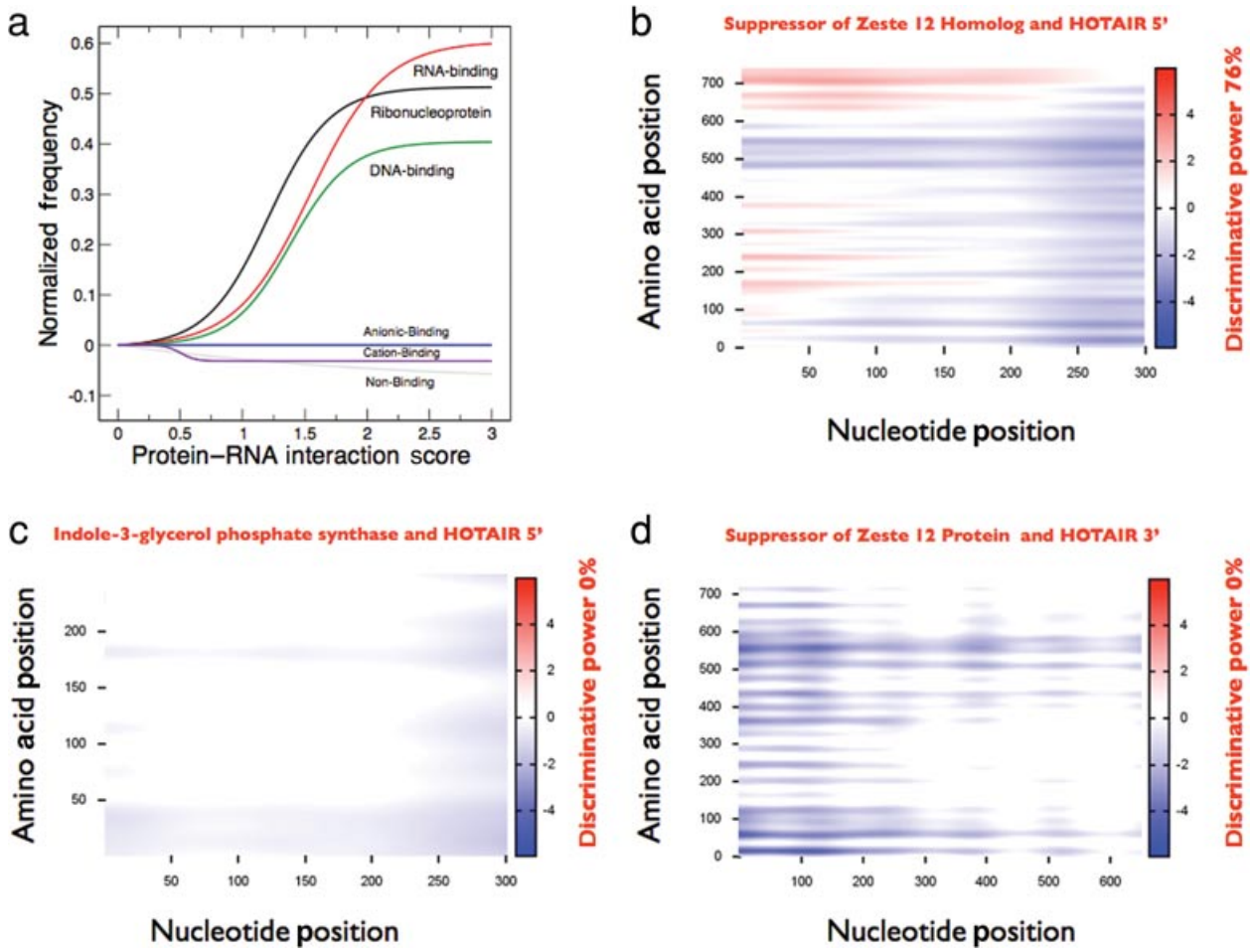


**FIGURE 1** | Predictions of RNA associations with protein classes. The catRAPID method is used to calculate the ability of proteins to interact with ncRNAs.[13] (a) RNA-binding, ribonucleo- and DNA-binding proteins show higher propensities to bind to RNA than anionic-, cation-, and non-nucleic-acid binding proteins.[12] (b) In *Homo sapiens,* The Suppressor of Zeste Homolog Suz12 is predicted to interact with HOTAIR 5′ (NCBI entry NR_003716.2 nucleotides 56–355; discriminative power 76%), which is in agreement with experimental evidence.[15] RNA-binding sites correspond to the C terminal VEFS-Box, which contains a Zinc finger domain, as well as amino acids 150–200 and 210–250, which are rich in positively charged residues; (c) indoleglycerol Phosphate Synthase (pdb code 1A53) belongs to the class non-nucleic-acid binding proteins (NNBP)[12] and shows negligible propensity to bind to HOTAIR 5′ (discriminative power 0%).[11,12] (d) Suz12 does not bind to HOTAIR 3′ (nucleotides 1553-2198; discriminative power 0%), as shown by experiments carried out in HeLa cells.[15]

1–10 proteins, respectively. In *Escherichia coli*, RNase P contains a large catalytic molecule of RNA (M1) and a small protein subunit (C5). Although there is evidence for a catalytic activity of the RNA alone *in vitro*,[15] C5 is strictly required for RNAse P function *in vivo*.[17] Binding of the protein subunit increases the affinity of the complex for the substrate up to 1000-fold. Indeed, C5 enhances substrate recognition and catalysis upon binding to the 5′ leader sequence of pre-tRNA.[18] In agreement with experimental evidence,[18] catRAPID predictions indicate that 92% of *E. coli* pre-tRNAs have strong propensity to interact with C5 (Figure 2).

### *The CSR Regulatory System*

The Csr system includes two critical components, CsrA and CsrB. CsrA is a 61-amino-acid protein related in sequence to several other RBPs.[20,21] CsrA is essential for the major decay pathway of transcripts and represents a model system to explore the prokaryotic RNA regulation.[22] The second regulatory component, CsrB, is a noncoding RNA molecule (fRNAdb code: FR283968), which forms a globular complex with approximately 18 CsrA polypeptides and antagonizes CsrA activity.[22] In agreement with previous observations,[20,21] catRAPID predictions indicate that CsrB contacts CsrA in multiple regions

  

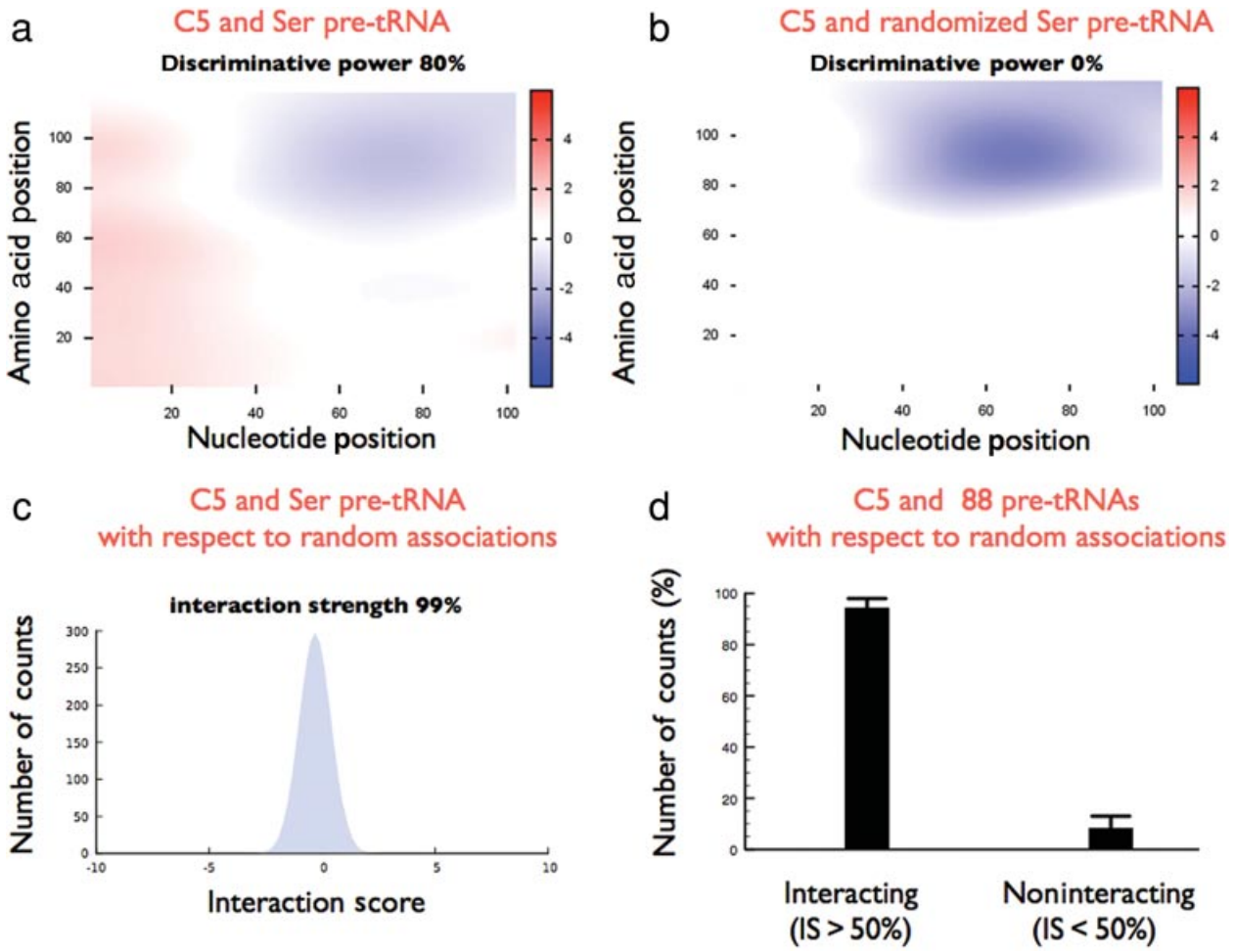


**FIGURE 2** | RNAse P. In *Escherichia coli*, RNase P is composed of a large catalytic molecule of RNA (M1 RNA) and a small protein subunit (C5 protein). This complex is required to process precursor tRNAs in functional tRNAs molecules. C5 protein enhances substrate recognition and helps catalysis by discriminating between substrate and product through the binding to the 5′ leader sequence of pre-tRNA.[18] The C5 RNA-binding domain spans the entire protein sequence. High interactions are predicted between C5 and several pre-tRNA molecules.[19] (a) High interaction propensity is predicted between C5 and Ser pre-tRNA (*E. coli* K12 2816667–2816575 nt; discriminative power 80%). The predicted binding site is located at the 5′ leader, in agreement with experimental evidence[19]; (b) randomization of Ser pre-tRNA sequence results in strong reduction of the C5 binding ability (discriminative power 0%); (c) C5 and Ser pre-tRNA are predicted to have higher interaction propensities than a random pool of $10^4$ associations with same protein and RNA lengths (interaction strength 99%, marked as blue area under the distribution curve). (d) We predict strong interaction propensities for all the pre-tRNA molecules reported to interact with C5.[19] More specifically, we find that 80 pre-tRNAs (i.e., 92% of the RNA set) have interaction strengths (IS) $>$ 50% with average interaction strength $=$ 82%.

at the 5′ and central region of the transcript, in correspondence to the repeated motifs CAGGATG, CAGGAAG, AAGGAAA, and AGGGAT[23] (Figure 3A). The ability of CsrB to sequester and antagonize the mRNA decay factor defines an important biological function for RNA.[20,21] catRAPID predicts very strong propensity for the CsrA-CsrB system (Figure 3B).

### *Comparisons with Other Methods*

The RPISeq predicts ribonucleoprotein interactions using the information contained in protein and RNA sequence patterns (http://pridb.gdcb.iastate.edu/RPISeq; see section *RPISeq*).[24] In agreement with catRAPID predictions and experimental evidence,[13,15] RPISeq predicts that HOTAIR 5′ binds to Suz12 with probabilities $P = 0.6$ (RPISeq-RF method) and 0.52 (RPISeq-SVM method). By contrast, Suz12 and HOTAIR 3′ are predicted to interact with $P = 0.65$ (RF) and 0.95 (SVM). As for the associations between C5 and Ser pre-tRNA, RPISeq reports $P = 0.35$ (RF) and 0.87 (SVM), whereas the CsrA-CsrB association is predicted with $P = 0.71$ (SVM) and 0.13 (RF). The negative controls indoleglycerol

  

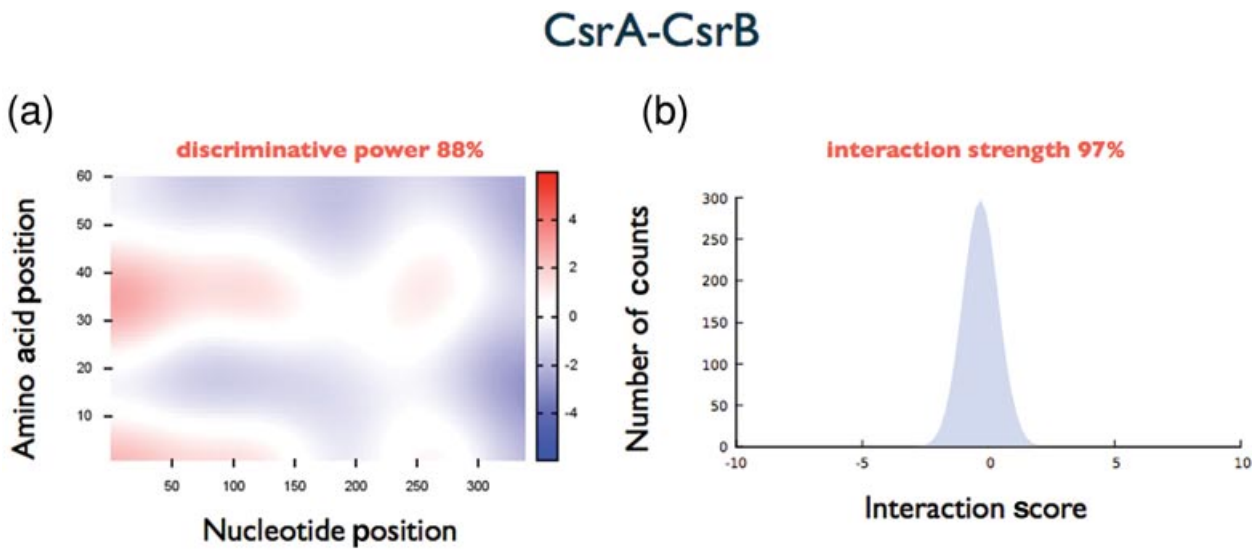


**FIGURE 3** | The CsrA-CsrB system. The central component of the calcium storage regulator system, CsrA, is a 61-amino-acid RNA-binding protein. This small protein, whose RNA-binding domain spans the entire polypeptide sequence, inhibits glycogen biosynthesis and catabolism, gluconeogenesis, and biofilm formation, whereas it activates glycolysis, acetate metabolism, motility, and flagellum biosynthesis.[20,21] A second component of the CSR system is untranslated CsrB RNA (fRNAdb code: FR283968), which binds to a number CsrA subunits, forming a large globular ribonucleoprotein complex (*Erwinia carotovora*). (a) In agreement with previous observations, CsrB is predicted to contact CsrA (discriminative power 88%)[13] and shows binding regions in the 5′ and central region of the transcript in correspondence to the repeated motifs CAGGATG, CAGGAAG, AAGGAAA, and AGGGAT.[23] (b) Very strong interaction is predicted for the CsrA–CsrB system with respect to a random pool of $10^4$ protein–RNA associations (interaction strength 97%, marked as blue area under the distribution curve).

phosphate synthase and randomized pre-tRNA are predicted to bind to HOTAIR 5′ and C5 with $P = 0.75$ (RF) / 0.64 (SVM) and 0.55 (RF) / 0.92 (SVM), respectively.

## Details about the catRAPID Method

### *The Interaction Profile*

The contributions of secondary structure, hydrogen bonding and van der Waals' are combined together into the *interaction profile* [see Eq. (1)]:

$$\underline{\Phi}_x = \alpha_S \underline{S}_x + \alpha_H \underline{H}_x + \alpha_W \underline{W}_x. \quad (1)$$

In Eq. (1), $\underline{X}$ indicates the physicochemical profile of a property $X$ calculated for each amino acid (nucleotide) starting from the $N$-terminus (5′). For example, the hydrogen bonding profile, denoted by $\underline{H}$, is the hydrogen bonding ability of each amino acid (nucleotide) in the sequence (Eq. 2):

$$\underline{H} = H_1, H_2, \ldots H_{\text{length}}. \quad (2)$$

Similarly, $\underline{S}$ represents the secondary structure occupancy profile and $\underline{W}$ the van der Waals' profile. The variable $x$ indicates RNA ($x = r$) or protein ($x = p$) profiles. Secondary structure, hydrogen bonding, and van der Walls contributions are calculated as follows.

### *Secondary Structure Propensities*

The secondary structure of a RNA molecule is predicted from its nucleotide sequence using the Vienna package.[25] Although the average predictive power of the RNAfold algorithm is roughly 70%, lower performances are expected for long noncoding RNAs because these transcripts are poorly characterized. To increase the amount of information that can be extracted from secondary structure predictions, ensembles produced with the RNAsubopt algorithm from Vienna suite were generated (http://www.tbi.univie.ac.at/~ivo/RNA/). The sampling of structures was performed with probabilities estimated through Boltzmann weighting and stochastic backtracking in the partition function. Six model structures, ranked by energy, are used as input for *cat*RAPID. For each model structure, the RNAplot algorithm was employed to generate secondary structure coordinates. Using the coordinates the *secondary structure occupancy* was defined by counting the number of contacts made by the nucleotide chain. High values of secondary structure occupancy indicate that base pairing occurs in regions with high propensity to form hairpin-loops, whereas low

  

values are associated with junctions or multiloops. The secondary structure of proteins was taken into account by calculating the Chou-Fasman[26] and Deleage-Roux[27] propensities for turn, $\beta$-strand and $\alpha$-helical elements. As the average predictive power of these models is around 60%, the individual propensities were combined to have better perfomances. The correlation between interaction propensities and secondary structure contributions is 0.73.

### *Hydrogen-Bonding Propensities*

The structural information on purine and pyrimidine contacts was extracted from a set of 41 nonredundant protein–RNA complexes.[28] Both the number and the frequency of hydrogen-bond contacts are used in the method. With respect to proteins, Grantham's[29] and Zimmerman's[30] scales were employed to estimate the propensity of amino acids to form hydrogen bonds. Other propensity scales were disregarded because they showed lower predictive power. The correlation between interaction propensities and hydrogen bonding contributions is 0.58.

### *Van der Waals' Propensities*

Similarly to hydrogen-bonding propensities, the information on purine and pyrimidine contacts was taken from a set of protein–RNA complexes.[28] Both the number and the frequency of van der Waals' contacts were used in *cat*RAPID. With respect to proteins, Kyte–Dolittle[31] and Bull-Breese[32] scales were employed to estimate the propensity to form van der Waals' contacts. Other propensity scales were disregarded because they showed lower predictive power. The correlation between interaction propensities and Van der Waals' contributions is 0.26 (estimated with a 10-fold cross-validation).

### *Interaction Propensity*

A discrete Fourier transform is employed to compare interaction profiles of different lengths:

$$\Psi_{k,x} = \sqrt{\frac{2}{\text{length}}} \sum_{n=0}^{\text{length}} \Phi_{n,x} \cos \times \left[ \frac{\pi}{\text{length}} \left( n + \frac{1}{2} \right) \left( k + \frac{1}{2} \right) \right] \quad k = 0, 1, \ldots \ell. \tag{3}$$

The number of plane waves is $\ell = 50$.

The *interaction propensity* $\pi$ is defined as the inner product between the protein propensity profile $\underline{\Psi}_p$ and the RNA propensity profile $\underline{\Psi}_r$ weighted by the *interaction matrix* I [Eq. (4)]:

$$\pi = \underline{\Psi}_p \cdot (\mathrm{I}\underline{\Psi}_r). \tag{4}$$

The interaction matrix I as well as the parameters $\alpha_S$, $\alpha_H$, and $\alpha_W$ are derived under the condition that the interaction propensities $\pi$ take maximal values for associations in the positive training set (and minimal or associations in the negative training set) [Eq. (5)]:

$$\mathrm{I}: \begin{cases} \max \underline{\Psi}_p \cdot (\mathrm{I}\underline{\Psi}_r) \, \forall \{\mathrm{r}, \mathrm{p}\} \in \{\text{positive training set}\}, \\ \min \underline{\Psi}_p \cdot (\mathrm{I}\underline{\Psi}_r) \, \forall \{\mathrm{r}, \mathrm{p}\} \in \{\text{negative training set}\}. \end{cases} \tag{5}$$

### *Training Set*

The structural data set (X-ray- and NMR-based models) was retrieved in March 2010 and consisted of 858 RNA–protein complexes available at the RCSB databank (http://www.pdb.org/). A cutoff of 7 Å for physical contacts was employed to discriminate between interacting and noninteracting protein–RNA pairs. The cutoff was selected in accordance with the average resolution of the structural complexes and led to define a positive dataset containing 7409 interacting protein–RNA pairs and a negative set containing 958 noninteracting protein–RNA pairs. The CD-HIT tool (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi) was used to filter out RNA and protein sequences with identities higher than 80% and 60%, respectively. After redundancy removal, the database contained 410 interacting ('positive set') and 182 noninteracting ('negative set') protein–RNA pairs. With regards to the composition of the positive set, protein–RNA associations were grouped into five functional classes: Ribosome and protein synthesis, splicing, transcription, tRNA synthesis, and viral RNA assemblies. For each class the interaction propensities of protein–RNA pairs are compared with the interaction propensities of the entire negative set. Performances were estimated with a 10-fold cross-validation. The algorithm to compute the interaction propensity with respect to the negative training set (discriminative power) is available at http://tartaglialab.crg.cat.

### *Scoring Function*

To evaluate the ability of *cat*RAPID to distinguish between interacting and noninteracting RNA–protein associations, the concept of discriminative power (dp) is introduced:

$$\mathrm{dp} = \frac{\sum_{i,n} \vartheta(\pi_i - \pi_n)}{\sum_{i,n} \vartheta(\pi_i - \pi_n) + \vartheta(\pi_n - \pi_i)}, \tag{6}$$

  

where $\pi_i$ indicates the interaction propensity of an interacting RNA–protein pair $i$, whereas $\pi_n$ represents the interaction propensity of noninteracting molecules $n$. The function $\vartheta(\pi_i - \pi_n)$ is 1 if $\pi_i - \pi_n > 0$ and 0 otherwise. Accordingly, the denominator equals the number of interacting pairs $I$ multiplied the number of noninteracting pairs $N$ [Eq. (7)]:

$$\sum_{i,n} \vartheta(\pi_i - \pi_n) + \vartheta(\pi_n - \pi_i) = I + N. \quad (7)$$

Following the definition given in Eq. (6), the discriminative power ranges from 0 to 1. The significance of predictions was evaluated by calculating $P$ values with ANOVA (two-tails' $t$-test). With regards to catRAPID's performances, the discriminative power associated with the nonredundant training dataset is 78%. The discriminative power associated with the redundant training dataset (X-ray and NMR structural models) is 90%.

### *Test Sets*

The NNBP[12] database was employed to evaluate the ability of *cat*RAPID to identify proteins that have little propensity to interact with RNA molecules. A database of 246 proteins was combined with RNA sequences of the positive set to generate 2500 random associations. The discriminative power of the algorithm was evaluated by comparing the interaction propensities of the positive set with the interaction propensities of these random associations. The NPInter database (http://www.bioinfo.org.cn/NPInter/)[33] was used to evaluate the ability of the algorithm to predict interactions between proteins and long noncoding RNAs. RNA sequences were obtained from the fRNAdb database (http://www.ncrna.org/frnadb/). We note that only a portion of the NPinter database shows direct physical evidence for protein–RNA interactions. The discriminative power of the algorithm was evaluated by comparing the interaction propensities of the NPinter database with the interaction propensities of the negative set. The receiver operating characteristic (ROC) analysis was introduced to compare performances of the algorithm in the training and test sets. In the analysis, interacting and noninteracting pairs of molecules represented the test sets and the quality of predictions was evaluated by calculating true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Performances were assessed by plotting the true positive rate (Sensitivity) in function of the false positive rate (1-Specificity) for different cutoff points:

$$\text{Sensitivity} = \text{TN}/(\text{TN} + \text{FP}). \quad (8)$$

The areas under the ROC curve are 0.75 for the training set (0.94 on the redundant dataset), 0.92 for the NNBP set and 0.88 for the NPinter database.

### *Comparison with Random Sets*

For each ribonucleoprotein association, a reference set is generated using random protein and RNA sequences that have exactly the same lengths as the molecules under investigation. The reference set contains random interactions between polypeptide and nucleotide sequences and represents a negative control because little interaction propensities are expected from these associations. In the calculations, the interaction propensity $\pi$ of a protein–RNA association is compared with the interaction propensities $\tilde{\pi}$ of the reference set ($10^4$ protein–RNA pairs). Using the interaction propensity distribution of the reference set, the scores are compared:

$$\begin{cases} m = \dfrac{1}{L}\displaystyle\sum_{i=1}^{\Lambda} \tilde{\pi}_i, \\ s^2 = \dfrac{1}{L}\displaystyle\sum_{i=1}^{\Lambda} (\tilde{\pi}_i - m)^2, \end{cases} \quad (9)$$

where the number of interactions is $L = 10^4$. From the distribution of interaction propensities, the interaction strength, IS, is computed as

$$\text{IS} = P(\tilde{\pi} \leq \pi). \quad (10)$$

## SEQUENCE PATTERNS FOR PROTEIN–RNA INTERACTIONS (RPISeq, SCRPRED, PPRINT, and PRINTR)

Sequence patterns are particularly useful to identify binding motifs in protein and RNA molecules. The algorithms SCRPRED,[34] PPRINT,[35] and PRINTR[36] exploit evolutionary information on protein sequences, whereas RPISeq[37] uses a reduced alphabet to describe protein and RNA sequences. Machine learning methods, such as support vector machines (SVMs), neural networks (NN), and random forest (RF), are used to identify RBPs (SCRPRED,[34] PPRINT,[35] and PRINTR[36]) as well as protein–RNA couples (RPISeq[37]).

### RPISeq

The RPISeq method (http://pridb.gdcb.iastate.edu/RPISeq/) consists of two classifiers: RPISeq-SVM (SVM classifier) and RPISeq-RF (random forest classifier).[24] In both algorithms, 343 features are used

  

**TABLE 2** | The RPISeq Method

*Dataset*: Two nonredundant datasets of RNA–protein interacting pairs were extracted from a total 943 protein–RNA complexes using a cutoff of 8 Å. Two datasets obtained from RNA immunoaffinity purification and microarray experiments are employed to test the performances.

*Method*: Each RNA–protein pair is represented by a 599-feature vector, of which 343 are related to protein properties and 256 describe RNA characteristics. Proteins are encoded using the conjoint triad feature: the 20 amino acids are classified into 7 groups according to their dipole moments and the volume of their side chain. RNAs are encoded 4-mers describing nucleotide frequencies. The output of the support vector machine (SVM) is a binary label indicating whether the given RNA–protein pair interacts or not. The authors use the sequential minimal optimization implementation to train the SVM classifier.

**TABLE 3** | The SRC PRED Method

*Dataset*: One hundred and sixty clusters of protein chain sequences (25% identity) retrieved from Protein Data Bank protein–RNA complexes with the highest number of dinucleotide RNA contacts (atom–atom distance <3.5 Å).

*Method I*: A vector representing global amino acid composition (GAC) of each protein sequence was concatenated to the evolutionary profile of each residue, represented as position-specific scoring matrices (PSSMs) computed by using the PSI-BLAST. Sliding windows of sizes ranging from 1 to 8 sequence neighbors were centered on each residue. Different combinations of feature matrices were tested and the best performing combination was retained. Different network architectures were optimized for every feature matrix varying the number of nodes in the hidden layer. Neural networks were trained to return a vector of 16 dinucleotide-binding prediction scores for a protein residue between 0 and 1, which are transformed to binary states of binding or nonbinding by choosing a cutoff.

*Method II*: Combinations of protein–RNA fragment pairs were generated matching each protein fragment to all RNA fragments. A fragment pair was labeled as '1' if any of the atoms from the protein fragment is in contact with any of the atoms of the RNA fragment at a threshold of 3.5 Å distance ('0' otherwise). The protein fragments are encoded in feature vectors built by concatenation of dinucleotide scores of each residue in a given fragment, while the RNA sequence fragments were encoded using dinucleotide compositions. The protein and RNA feature vectors for each fragment pair in a given complex are concatenated forming a fragment-pair feature matrix per complex. The neural network predicts a one-dimensional binary vector encoding the 'paired' or 'nonpaired' state of the protein–RNA fragment pairs.

to encode the protein sequence and 256 features are used to encode the RNA sequence. Proteins are encoded using the conjoint triad feature representation: the 20 amino acids are classified into 7 groups: {A, G, V}, {I,L,F,P},{Y,M,T,S},{H,N,Q,W},{R,K},{D,E}, and {C}. Thus, each protein sequence is represented by a $7 \times 7 \times 7$ or 343-dimensional vector, where each element of the vector corresponds to the normalized frequency of the corresponding 3-mer in the sequence. Based on the k-mer frequency representation of RNA sequences, RNA sequences were encoded using a $4 \times 4 \times 4$ or 256-dimensional vector, in which each feature represents the 4-mer normalized frequency appearing in a RNA sequence (e.g., AUUG, CCAU, GACA). On two nonredundant benchmark datasets extracted from the Protein–RNA Interface Database (PRIDB), the method performs with accuracies varying between 76–90% (RF) and 73–87% (SVM) and areas under the ROC curves of 0.92–0.97 (RF) and 0.81–0.85 (SVM). RPISeq classifiers trained on PRIDB predict a number of noncoding RNA–protein interactions present in the NPInter database[33] (Table 2; see also catRAPID test sets).

## SRCPRED

SRCPRED (http://tardis.nibio.go.jp/netasa/srcpred/) uses neural networks to predict RNA-interacting amino acids by means of sequence (global amino acid composition, GAC scores) and evolutionary information (position-specific scoring matrix, PSSM scores).[34] To train the neural networks, the authors employed protein sequence-neighbors feature matrices, and protein–RNA fragment-pair feature matrices. The area under the ROC curve ranges from 0.61 to 0.84 for four tested RNA functional classes (viral RNA, mRNA, tRNA, rRNA). The analysis of PSSM scores[38] indicates that, although more complex and varied than protein–DNA interactions, the substitution patterns of the functional residues in RBPs are significantly constrained during evolution (Table 3).

## PPRINT

Combining evolutionary information and SVMs, PPRINT was developed to predict RNA-binding sites in protein sequences (http://www.imtech.res.in/raghava/pprint/).[35] The original dataset contains 86 RNA interacting protein chains extracted from structures of RNA–protein complexes. Using

  

**TABLE 4** | The PPRINT Method

*Dataset*: Eighty-six RNA interacting protein chains extracted from Protein Data Bank.
*Method*: A support vector machine trained on protein sequence patterns. The performance is high when multiple sequence alignments are used in the form of PSSM profiles (Matthew's correlation coefficient of 0.45).

PSI-BLAST,[38] nonredundant protein chains are selected for training (sequence similarity <70%). The evolutionary information is derived from PSSM generated during PSI-BLAST searches against a nonredundant database of protein sequences. To train PPRINT, the authors use fixed-length patterns of amino acids. A pattern is considered positive when the central residue is in direct contact with RNA (negative otherwise). The patterns are converted into binary vectors and each amino acid was described using a vector of 20 entries. PPRINT performs with an accuracy of 76% in predicting residues that contact RNA (Matthew's correlation coefficient of 0.45; Table 4).[35]

## PRINTR

PRINTR (http://210.42.106.80/printr/) uses SVMs and PSSMs to predict protein residues at RNA-binding surfaces.[36] The method achieves an area under the ROC curve of 0.83 using information derived from multiple sequence alignment, solvent accessibility, and secondary structure. The stringent criterion used to choose nonhomologues dataset (<30% sequence homology) and the combination of SVM and PSSMs are key features of the method (Table 5).

## STRUCTURE-BASED PREDICTIONS OF RNA-BINDING SITES (RNABindR, Struct-NB, PRIP, PatchFinderPlus, SPOT, AND OPRA)

The algorithms Struct-NB,[39] PRIP,[40] PatchFinderPlus,[41] SPOT,[42] and OPRA[43] predict RNA-binding using properties of protein surfaces. SVM and Naïve Bayes Classifiers (NBCs) trained on structural data are employed to analyze surface features. The RNABindR method combines structural information with sequence-based predictions of hydrophobicity and entropy.[37] In general, the success of structure-based predictive methods can provide great structural detail on substrate-binding clefts, but is greatly limited by the availability of protein–RNA complexes as templates.[42]

**TABLE 5** | The PRINTR Method

*Dataset*: One hundred and nine protein chains from Terribilini et al.[37] database.
*Method*: After position-specific scoring matrices (PSSMs) generation, five different encoding schemes are employed: (1) single sequence; (2) multiple sequence alignment; (3) single sequence plus predicted secondary structure; (4) multiple sequence alignment plus predicted secondary structure; (5) multiple sequence alignment plus secondary structure and solvent accessibility information. In the case of single sequence, the feature vector representing a residue is calculated using an optimized sliding window. In case of multiple alignments, the feature vector representing a residue is extracted using PSSM. Moreover, DSSP (http://swift.cmbi.ru.nl/gv/dssp/) is used to compute residue solvent-accessible surface area, and secondary structure information was encoded as helix, strand, and coil.

## RNABindR

RNaBindR (http://einstein.cs.iastate.edu/RNABindR/) is a classifier that predicts RNA-binding regions.[37] The variables used in RNaBindR method are relative accessible surface area (rASA), sequence entropy, hydrophobicity, secondary structure, and electrostatics. The rASA is computed using the program Naccess (http://wolf.bms.umist.ac.uk/naccess/). Sequence entropy is estimated using the relative entropy for each residue from the HSSP database (http://www.cmbi.kun.nl/gv/hssp/). Hydrophobicity of each residue is obtained from the consensus normalized hydrophobicity scale derived by Sweet and Eisenberg.[44] In addition, the authors use information on secondary structures extracted from the protein databank. Electrostatic potentials are calculated using the program APBS (http://agave.wustl.edu/apbs/). In the leave-one-out cross-validation procedure, RNABindR identifies interface residues with 85% accuracy. The training set is extracted from structures of known RNA–protein complexes solved by X-ray crystallography. Proteins with >30% sequence identity or structures with resolution worse than 3.5 Å are removed using PISCES.[45] Amino acids in the RNA–protein interface are identified using ENTANGLE.[46] Positively charged amino acids arginine and lysine show the highest interface propensities, consistently with their enhanced ability to participate in interactions with bases and negatively charged phosphate backbone. Another favored residue, histidine, is found to participate in stacking interactions with

**TABLE 6** | The RNABindR Method

*Dataset*: A set of 109 nonredundant protein chains containing a total of 25,118 amino acids.
*Method*: A Naive Bayes classifier trained using data from 108 chains and evaluated on the 109th chain. The algorithm employs relative accessible surface area, sequence entropy, hydrophobicity, secondary structure, and electrostatic potential (calculated on known structures).

RNA bases through the imidazole ring. In contrast, phenylalanine and negatively charged amino acids glutamate and aspartate are significantly underrepresented in interfaces. Hydrophobic residues such as leucine, isoleucine, valine, and alanine are significantly depleted at interfaces (Table 6).

## Struct-NB

Struct-NB (http://www.public.iastate.edu/~ftowfic) uses an ensemble of NBCs and a structural-based Gaussian Naïve Bayes classifier (GNBC) to predict RNA-binding sites.[39] The NBC is trained on sequence-based features present in the RNABindR algorihm,[37] whereas the structural-based GNBC exploits two main structural features: the surface roughness (i.e., degree of irregularity on the surface) and the CX value (i.e., the ratio of the volume of atoms that occupy a 6 Å sphere compared to the empty volume within the sphere). Struct-NB achieves an area under the ROC curve of 0.75. With respect to other sequence-based classifiers, the classification performance is improved with integrated structural information. The analysis of the structural features reveals that protein–RNA interface residues are associated with higher degree of irregularity at the surface (amino acids protruding out of the surface) compared to noninterface residues (Table 7).

**TABLE 7** | The Struct-NB Method

*Dataset*: One hundred and forty-seven protein structures from Terribilini et al.[37] database.
*Method*: The sequence-based NBC is trained to label the target residue with a '+' (protein–RNA interface residue) or a '−' (otherwise), using sequence-based features of the RNABindR method (Terribilini et al.[37]). The structural information of a residue is encoded to attribute values within a sequence window. Two structural properties of amino acid residues are employed: surface roughness (i.e., degree of irregularity of that point at the surface) and CX value (i.e., the ratio of the volume of atoms that occupy a 6 Å sphere compared to the empty volume within the sphere). The structural-based Gaussian Naïve Bayes classifier is similar to the NB classifier, except that the attribute values are numerical. Overall, 5 NB classifiers were trained, one for each feature representation: sequence-based, Calpha-based (CX score for the alpha carbon atom as the score for the corresponding residue), average CX-based (average of the CX scores for all atoms), R group-based (average of the CX scores atoms in the R-group only), and roughness-based feature representations, respectively. Struct-NB returns the interface propensity of a property as a measure of the preference for a value (or a range of values) of a property among the interface residues (relative to the entire set of surface residues).

**TABLE 8** | The PRIP Method

*Dataset*: One hundred and forty-four protein chains from Terribilini et al. (2007) database.
*Method*: A residue is classified as interacting or noninteracting based on the features of three types of interface residues (cutoff of 5 Å): (1) sequential patch (or sequence sliding windows) of size *n*, i.e., the *n* residues nearest to the residue (center residue); (2) spatial patch of size *n*, i.e. the set of the *n* residues with the smallest euclidean distance between their Calpha atoms and the Calpha atom of the center residue; (3) topological patch, i.e., the *n* vertices with the smallest geodesic distances (shortest paths) to the center vertex. Features such as amino acids indexes, sequence profiles, accessible surface area, betweenness centrality, and retention coefficient are compared with respect to their predictive power to detect interface residues.

## PRIP

PRIP (http://qfab.imb.uq.edu.au/PRIP) exploits NBCs and SVMs combined with graph-theoretic properties of interface residues.[40] Contact graphs derived from interface residues neighborhood patches (sequential and spatial) are used to derive the model. Moreover, topological features, such as betweenness centrality, are calculated based on the entire contact map of a protein chain and define clusters of residues (close in space and sequence) present in the patch. Overall, SVMs outperform NBCs by about 5%, and spatial patches give better signal than topological and sequential patches (Table 8). The area under the ROC curve is 0.83 and is calculated using spatial patch, ASA, betweenness centrality, and retention coefficient (Table 8). This analysis indicates that network theory can describe the collective properties of interface residues accounting for their binding affinity and specificity.

## PatchFinderPlus

Using an ensemble of protein features and specific properties extracted from electrostatic patches with the algorithm PatchFinderPlus (http://pfp.technion.

  

**TABLE 9** | Method Based on PatchFinderPlus

| |
|---|
| *Dataset*: The set includes 76 nonredundant structures. As a control, the authors use a nonredundant database of 246 non-nucleic-acid-binding protein (NNBP) chains. |
| *Method*: The authors use a SVM classifier to distinguish between the nonredundant set of RNA-binding proteins (RBPs) and the NNBPs, as well as between the RBPs and the subset of NNBPs with large positive patches. For training, a feature vector is employed. The vector includes 40 sequence and structural parameters extracted from both the electrostatic patches and the whole protein. To test the method, the authors apply a cross-validation test, where for each SVM run, one protein is extracted from the training and tested separately. |

**TABLE 10** | The SPOT Method

| |
|---|
| *Dataset*: Seven datasets used: (1) 250 representative RNA-binding holo-domain (i.e., bound) structures (RB250 libary); (2) 6761 non-RNA-binding domain; (3) 212 nonredundant RNA-binding holo-domains structures; (4) 75 RNA-binding apo-domains (i.e., unbound) structures with 45–100% sequence identity to the 75 out of the 250 holo-domains; (5) 331 DNA-binding domains structures; (6) 292 RNA-binding domains belonging to five superfamilies (canonical, noncanonical, splicing factor U2AF subunits, Smg-4/UPF3, and GUCT); (7) 2076 domains from previously collected structural genomics targets. |
| *Method*: The target structure is scanned against templates with sequence identity $< 30\%$ in a reference library (RB250) using the structural alignment program TM-align. If the structural similarity score is higher than a threshold, the protein–RNA complex structure is predicted by replacing the template structure with the aligned target structure. Two structural similarity scores are employed: one is based on the raw TM-score and the other one is based on Z-score. If the lowest binding energy between the target protein and template RNA is lower than a threshold and the structure similarity is higher than a threshold, the target is predicted as an RBP and its RNA-binding site can be predicted from the predicted protein–RNA complex structure. If no matching template is found to satisfy these two thresholds, this target is predicted as a non-RNA-binding protein. |

ac.il/), the authors train a SVM to distinguish RBPs from other positively charged proteins that do not bind nucleic acids.[41] The method is applied on proteins possessing the RNA recognition motif (RRM) and successfully classified as RBPs from RRM domains involved in protein–protein interactions (Table 9). In addition to the features extracted from the surfaces patches alone, the authors calculate other global parameters of each protein, such as the molecular weight, surface accessibility, the size of the largest clefts, and the overlap between the clefts and the patches. Among the general properties, the molecular weight and surface accessibility are significantly lower in RBPs compared to NNBP.[41,47] The dipole and quadrupole moments for all the proteins in the datasets are also calculated. As expected, the dipole moment is significantly higher in the RBPs compared to NNBPs.

## SPOT

SPOT (http://sparks.informatics.iupui.edu) employs structural alignments of known protein–RNA complex structures and a statistical energy function to discriminate RBPs from nonnucleic acid binding proteins.[42] The relative structural similarity measured with alignments shows a Matthew's correlation coefficient MCC of 0.48 and efficiently discriminate RBPs from nonbinding proteins (Table 10). Optimal performances are reached combining a number or scores such as the binding affinity, the raw structural alignment and a global Z-score to measure structural similarity. The method is applied to predict RNA-binding residues on a dataset of nonredundant RNA-binding domains (MCC of 0.72), holo-targets (MCC of 0.56), and apo-targets (MCC of 0.56). When applied to SCOP RNA-binding domain superfamily prediction, the method achieves a 86% success rate. Among 2076 structural genomics domains of unknown function, the authors observe that 80% of the predicted targets are putative RBPs according to NCBI annotations. This study reveals the importance of dividing structures into domains, using a Z-score to measure structural similarity, and a statistical energy function to measure protein–RNA-binding affinity. One advantage of the method is the simultaneous prediction of protein–RNA complex structures.

## OPRA

OPRA (Optimal Protein–RNA Area) was developed to identify RNA-binding sites on proteins surfaces.[43] For each protein residue, a predictive score is calculated using protein–RNA interface propensities weighted by ASA (Table 11). As individual propensities are found to be a poor indicator of protein–RNA interfaces (residues with high individual interface propensity scores are not necessarily involved in protein–RNA interactions), the authors employ patch energy values modified by scores of neighboring residues. Remarkably, 80% of the test set is correctly predicted to interact. This study suggests that protein–RNA-binding determinants are laying on the

  

**TABLE 11** | The OPRA Method

| |
|---|
| *Dataset*: One hundred and seventy filtered protein–RNA complexes from PDB, comprising 316 nonredundant protein–RNA interactions (cutoff $\leq$4 Å); 282 interactions were used for training and 33 for testing. |
| *Method*: Statistical interface propensities of residues/ribonucleotides converted to free-energy estimates (statistical potentials). Considering statistical potentials as additive by surface area units, the authors identified residues/ribonucleotides that have favorable tendency to be at the interface (effective surface-to-interface energy transfer). |

protein side, at least from a residue composition point of view.

## CONCLUSIONS

The number of reported protein–RNA complexes is rapidly increasing. This growth is visible in the PDB database as the yearly increase of deposited protein–nucleic acid complexes and in the PubMed database as the change in the number of publications associated with the term 'RNA-binding proteins'. Yet, the determination of structures for proteins in complex with their partner RNAs is laborious and slow and there is a large demand for the development of computational methods for predicting such structures either from homologue structures of the components or directly from sequences.

The critical assessment of prediction of interactions (CAPRI) experiment (http://www.ebi.ac.uk/msd-srv/capri), a blind international docking competition to evaluate performances of protein–protein computational docking methods, recently proposed the first protein–RNA complex as target.[48] The study of protein–RNA targets indicates the growing interest in computational prediction of ribonucleoprotein structures and consequently, the need for new methods to solve protein–RNA docking problems. The CAPRI experiment encourages modeling groups to adapt existing protein–protein docking methods or develop new ones for protein–RNA docking problems. Indeed, previous methods could be adapted to predict the three-dimensional structure of protein–RNA complexes[49] but, owing to the high flexibility of RNA molecules, generation a protein–RNA model from the unbound structure is highly challenging.[50] The importance and difficulties of RNA modeling have also motivated the recent CASP-like challenge *RNA-Puzzles* to predict three-dimensional structure of RNA.[51]

Despite the success of structural genomics efforts,[52] the number of solved protein–RNA structures substantially lags behind the number of possible protein–RNA complexes and is underrepresented with respect to many RNA-binding motifs. Because of the difficulties associated with the experimental determination of protein–RNA complexes and RNA-binding sites in proteins,[53,54] there is an urgent need for reliable computational methods. At present, 289 X-ray and 116 NMR protein–RNA complexes solved at $\sim$3 Å resolution constitute the principal source of information for structural models. Various methods for predicting RNA structures and RNP complexes based on low-resolution experimental could be developed.[55,56] For instance, the structure of many macromolecular complexes can be modeled by using cryo-EM maps and restraints from biochemical experiments and other bioinformatics-based predictions.[57] Yet, dedicated algorithms for automated predictions of protein–RNA interactions remain to be developed. As a matter of fact, more than 300 intrinsically disordered proteins have been recently classified as RNA binding.[58] Given the difficulty of assigning reference structures to natively unfolded proteins, it is likely that new predictors of protein-RNA associations should be based on primary rather than tertiary structure.

Synergy is expected from the combination of predictive methods with low-resolution experimental analyses. Structural probing experiments such as footprinting and crosslinking can provide information about RNA secondary structure, inter- and intramolecular interactions,[59] whereas SAXS and cryo-EM experiments can be used to obtain information about the shape of macromolecular complexes.[60] In particular, through-space distance constraints derived from biochemical experiments could provide crucial information to determine RNA folding. Small number of long-range, through-space distance constraints are sufficient to limit the conformational space enough to allow accurate structure predictions.[61] Indeed, experimental methods such as site-directed hydroxyl radical footprinting, cross-linking and fluorescence resonance energy transfer will provide valuable information to build new models for ribonucleoprotein associations.[61]

  

## ACKNOWLEDGMENTS

The authors would like to thank Domenica Marchese, Dr Pia Cosma and Dr R. Johnson for stimulating discussions. Our work was supported by Spanish Ministry of Economy and Competitiveness.

## REFERENCES

1. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. *Science* 2005, 309:1559–1563.
2. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007, 447:799–816.
3. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 2006, 7(suppl 1):S4.1–9.
4. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 2007, 316:1484–1488.
5. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. *Mol Cell* 2011, 43:904–914.
6. Khalil AM, Rinn JL. RNA-protein interactions in human health and disease. *Semin Cell Dev Biol* 2011, 22:359–365.
7. Chen Y, Varani G. Protein families and RNA recognition. *FEBS J* 2005, 272:2088–2097.
8. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 2008, 582:1977–1986.
9. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res* 2002, 30:1427–1464.
10. Cammer S, Carter CW Jr. Six Rossmannoid folds, including the Class I aminoacyl-tRNA synthetases, share a partial core with the anti-codon-binding domain of a Class II aminoacyl-tRNA synthetase. *Bioinformatics* 2010, 26:709–714.
11. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 2011, 40:D302–D305.
12. Stawiski EW, Gregoret LM, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure. *J Mol Biol* 2003, 326:1065–1079.
13. Bellucci M, Agostini F, Masin M, Tartaglia GG. Predicting protein associations with long noncoding RNAs. *Nat Methods* 2011, 8:444–445.
14. Goodrich JA, Kugel JF. Non-coding-RNA regulators of RNA polymerase II transcription. *Nat Rev Mol Cell Biol* 2006, 7:612–616.
15. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010, 329:689–693.
16. Lundblad EW, Altman S. Inhibition of gene expression by RNase P. *N Biotechnol* 2010, 27:212–221.
17. Sakano H, Yamada S, Ikemura T, Shimura Y, Ozeki H. Temperature sensitive mutants of *Escherichia coli* for tRNA synthesis. *Nucleic Acids Res* 1974, 1:355–371.
18. Koutmou KS, Zahler NH, Kurz JC, Campbell FE, Harris ME, Fierke CA. Protein-precursor tRNA contact leads to sequence-specific recognition of 5′ leaders by bacterial ribonuclease P. *J Mol Biol* 2010, 396:195–208.
19. Koutmou KS, Zahler NH, Kurz JC, Campbell FE, Harris ME, Fierke CA. Protein-precursor tRNA contact leads to sequence-specific recognition of 5′ leaders by bacterial ribonuclease P. *J Mol Biol* 2010, 396:195–208.
20. Romeo T, Gong M. Genetic and physical mapping of the regulatory gene csrA on the *Escherichia coli* K-12 chromosome. *J Bacteriol* 1993, 175:5740–5741.
21. Liu MY, Yang H, Romeo T. The product of the pleiotropic *Escherichia coli* gene csrA modulates glycogen biosynthesis via effects on mRNA stability. *J Bacteriol* 1995, 177:2663–2672.
22. Liu MY, Romeo T. The global regulator CsrA of *Escherichia coli* is a specific mRNA-binding protein. *J Bacteriol* 1997, 179:4639–4642.
23. Liu MY, Gui G, Wei B, Preston JF 3rd, Oakford L, Yüksel U, Giedroc DP, Romeo T. The RNA molecule CsrB binds to the global regulatory protein CsrA and

  

antagonizes its activity in *Escherichia coli*. *J Biol Chem* 1997, 272:17502–17510.

24. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA–protein interactions using only sequence information. *BMC Bioinformatics* 2011, 12:489.
25. Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. The Vienna RNA Websuite. *Nucleic Acids Res* 2008, 36:W70–W74.
26. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* 1978, 47:45–148.
27. Deléage G, Roux B. An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng* 1987, 1:289–294.
28. 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, 22:2746–2752.
29. Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974, 185:862–864.
30. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* 1968, 21, 170–201.
31. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, 157:105–132.
32. Bull HB, Breese K. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Arch Biochem Biophys* 1974, 161:665–670.
33. Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, Chen L, Lu H, Zhao Y, Chen R. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res* 2006, 34:D150–D152.
34. Fernandez M, Kumagai Y, Standley DM, Sarai A, Mizuguchi K, Ahmad S. Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC Bioinformatics* 2011, 12(suppl 13):S5.
35. Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit* 2011, 24:303–313.
36. 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.
37. Terribilini M, Sander JD, Lee J-H, Zaback P, Jernigan RL, Honavar V, Dobbs D. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res* 2007, 35:W578–W584.
38. Liu T, Geng X, Zheng X, Li R, Wang J. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. *Amino Acids* 2012, 42:2243–2249.
39. 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.
40. Maetschke SR, Yuan Z. Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *BMC Bioinformatics* 2009, 10:341.
41. Shazman S, Mandel-Gutfreund Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput Biol* 2008, 4:e1000146.
42. Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol* 2011, 8:988–996.
43. Pérez-Cano L, Fernández-Recio J. Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins* 2010, 78:25–35.
44. Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J Mol Biol* 1983, 171:479–488.
45. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003, 19:1589–1591.
46. Allers J, Shamoo Y. Structure-based analysis of protein–RNA interactions using the program ENTANGLE. *J Mol Biol* 2001, 311:75–86.
47. Hobohm U, Sander C. Enlarged representative set of protein structures. *Protein Sci* 1994, 3:522–524.
48. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. *Proteins* 2010, 78:3073–3084.
49. Cheng TM-K, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins* 2007, 68:503–515.
50. Gherghe CM, Leonard CW, Ding F, Dokholyan NV, Weeks KM. Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *J Am Chem Soc* 2009, 131:2541–2546.
51. Cruz JA, Blanchet M-F, Boniecki M, Bujnicki JM, Chen S-J, Cao S, Das R, et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* 2012.
52. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, et al. PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp* 2010.
53. Ke A, Doudna JA. Crystallization of RNA and RNA-protein complexes. *Methods* 2004, 34:408–414.
54. Wu H, Finger LD, Feigon J. Structure determination of protein/RNA complexes by NMR. *Meth Enzymol* 2005, 394:525–545.

  

55. Das R, Kudaravalli M, Jonikas M, Laederach A, Fong R, Schwans JP, Baker D, Piccirilli JA, Altman RB, Herschlag D. Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci USA* 2008, 105:4144–4149.
56. Mertens HDT, Svergun DI. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J Struct Biol* 2010, 172:128–141.
57. Jurica MS. Detailed close-ups and the big picture of spliceosomes. *Curr Opin Struct Biol* 2008, 18:315–320.
58. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 2012, 149:1393–1406.
59. Weeks KM. Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol* 2010, 20:295–304.
60. Lipfert J, Doniach S. Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annu Rev Biophys Biomol Struct* 2007, 36:307–327.
61. Ding F, Lavender CA, Weeks KM, Dokholyan NV. Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods* 2012.
62. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci* 2009, 106:11667–11672.
63. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* 1983, 35:849–857.

  