Computational and Structural Biotechnology Journal 23 (2024) 4036–4046

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

RNA-protein interaction prediction without high-throughput data: An overview and benchmark of *in silico* tools

Sarah Krautwurst [a,b,*], Kevin Lamkiewicz [a,b,c,1]

[a] *RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany*
[b] *European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany*
[c] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstr. 4, 04103 Leipzig, Germany*

ARTICLE INFO

Keywords:
RNA-protein interactions
RNA-protein interaction prediction
Deep learning
In silico RPI prediction
CLIP-Seq

ABSTRACT

RNA-protein interactions (RPIs) are crucial for accurately operating various processes in and between organisms across kingdoms of life. Mutual detection of RPI partner molecules depends on distinct sequential, structural, or thermodynamic features, which can be determined via experimental and bioinformatic methods. Still, the underlying molecular mechanisms of many RPIs are poorly understood. It is further hypothesized that many RPIs are not even described yet. Computational RPI prediction is continuously challenged by the lack of data and detailed research of very specific examples. With the discovery of novel RPI complexes in all kingdoms of life, adaptations of existing RPI prediction methods are necessary. Continuously improving computational RPI prediction is key in advancing the understanding of RPIs in detail and supplementing experimental RPI determination. The growing amount of data covering more species and detailed mechanisms support the accuracy of prediction tools, which in turn support specific experimental research on RPIs. Here, we give an overview of RPI prediction tools that do not use high-throughput data as the user's input. We review the tools according to their input, usability, and output. We then apply the tools to known RPI examples across different kingdoms of life. Our comparison shows that the investigated prediction tools do not favor a certain species and equip the user with results varying in degree of information, from an overall RPI score to detailed interacting residues. Furthermore, we provide a guide tree to assist users which RPI prediction tool is appropriate for their available input data and desired output.

1. Introduction

Interactions between RNAs and proteins (RPIs) are essential for multiple molecular processes in biological entities. Key players for RPIs are RNA-binding proteins (RBPs), which are involved in gene expression, RNA processing, modification, and degradation of RNA [1–3]. Interaction of mRNAs with RBPs can initiate or regulate protein synthesis [4–6]. The function of microRNAs (miRNAs) (for the regulation of gene expression) depends on RBPs [7,8], same as for long non-coding RNAs (lncRNAs) [9,10]. Due to the vast variety of processes in which RBPs are involved, their functionalities are also linked to diseases [11–13].

Most knowledge on RBPs originates from eukaryotic systems, especially the human organism [14], for which around 1,500 RBPs are annotated [15]. However, data for bacteria and viruses is sparse, as for a typical bacterium, around 180 RBPs are known [16]. For viruses, most RPI research focuses on host-virus rather than intra-viral RPIs, which started to set off only a decade ago [17].

RPIs are realized dynamically with a one-sided or mutual conformational change of the RNA and protein partner [18–20]. Additionally to this conformational change, many residues may not be part of the interaction directly but are still crucial for binding site flexibility and correctly positioning functional residues in RPIs [21,22]. The interacting protein binds RNA molecules either specifically (e.g. based on RNA modifications [23], sequential or structural motifs [24,25]) or non-specifically (e.g. dsRNA or ssRNA in general [24,26,27]). Within the protein, aromatic and positively charged amino acids are often involved in contacting the RNA partner, especially since they can form specific and strong interactions like salt bridges and π-stacking with the nucleobases [24,28]. However, the backbone of the RNA is associated more often with the protein than the bases in RPIs [29]. Solvent ac-

[*] Corresponding author at: RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany.
E-mail address: sarah.krautwurst@uni-jena.de (S. Krautwurst).

[1] Currently without academic affiliation.

https://doi.org/10.1016/j.csbj.2024.11.015
Received 20 August 2024; Received in revised form 5 November 2024; Accepted 5 November 2024
Available online 8 November 2024

cessibility and structural positions of the amino acids are decisive for interaction as well [21]. Usually, RNA binding domains are the main interaction area. Single amino acids outside of such domains can take further action toward contacting the RNA [1,28].

Experimental detection (*in vitro* and *in vivo*) of RPIs [30,31] can focus on an RNA molecule of interest to characterize potential proteins binding to it (e.g. RAT/TRAP, RNA affinity in tandem / tagged RNA affinity purification [32,33]) or on a known RBP to identify interacting RNAs (RIP-Chip, RNA immunoprecipitation chip [34]; CLIP, cross-linking immunoprecipitation [35]). When implemented accurately, *in vitro* methods can effectively predict *in vivo* RPIs and expand the landscape of RPI knowledge, e.g., by distinguishing binding specificity and defining context [36–38]. Still, *in vitro* methods may result in RPIs that are not physiologically relevant [1], which can be complemented by utilizing *in vivo* approaches. Non-crosslinking *in vivo* methods like RIP-Chip [34], which combines immunoprecipitation with RT-PCR and microarrays, can come along with noise issues, co-immunoprecipitation of unwanted additional proteins, false positives from re-associated proteins and RNAs after cell lysis [39,40], or no possibility to identify the binding site specifically because of mild conditions for preserving the non-covalent RPIs [41]. Instead of microarrays, RIP-Seq combines RNA immunoprecipitation with high-throughput sequencing [42]. Alternatives are CLIP-based (crosslinking and immunoprecipitation) methods (e.g. HITS-CLIP, iCLIP, Par-CLIP), which solve some of the former issues and carry other disadvantages. For example, HITS-CLIP (high-throughput sequencing CLIP) enables large-scale RPI detection [43], but the UV radiation from the UV-crosslinking step prior to immunoprecipitation may lead to mutagenesis [44], or specificity issues during crosslinking (biased for ssRNA, pyrimidines, certain amino acids) [41,45]. Par-CLIP (Photoactivatable-Ribonucleoside-Enhanced CLIP) [46] and iCLIP (individual nucleotide resolution CLIP) [47] both provide specific resolutions of interaction-involved nucleotides and binding sites but require long technical procedures (Par-CLIP [48]) and demanding set-ups (iCLIP [49]), involving numerous reaction steps. In general, *in vivo* techniques grant biologically more specific (e.g. regarding tissue/cell lines) interaction data, which is beneficial for specific training of computational RPI prediction algorithms. The potentially high false negative [50,51] and false positive [52,53] rate of these experiments however could lead to biases in training of such algorithms.

The above problems of experimental RPI detection are continuously being tackled with modern and up-to-date experimental workflows, identifying and revising the validity and interpretation of the methods [54]. Still, they can be expensive, time-consuming, or require challenging set-ups [30]. Furthermore, meeting the exact conditions for the RPIs to happen and be detectable *in vivo* (e.g. tissue-specificity, cell-cycle specificity, time-dependency, robustness of interaction) can prove difficult [55]. Therefore, experimental approaches can be supplemented with bioinformatics methods. Algorithms help analyze existing experimental data sets in more detail [56,57], investigate available sequences and structures for motifs and properties [58], predict binding sites of RNA-protein partners based on similar known interactions [59], or classify whether an RPI is probable [60]. Input for the algorithms is mostly the sequence or PDB (Protein Data Bank) [61] 3D structure of either or both RPI partners. The output ranges widely between tools, e.g., getting an overall interaction score for an RNA and protein pairing [62,63], sequence motifs contributing to an RPI [56,64], a potential binding site area highlighted in a protein structure [65–67], or specific interacting residues between an RNA-protein complex [68].

Many RPI prediction tools are available, varying in aspects like background data, feature selection, machine-learning algorithm, or output extent. In recent years, several tools and workflows have been proposed that utilize experimental data from, e.g., CLIP-Seq experiments. While these resources are valuable and often used to determine specific binding motifs or residues of a protein of interest, there are scientific questions or use cases where such data is unavailable. For an overview of tools available that use high-throughput sequencing (HTS) data, we refer to [69–71].

Here, we present a comparative analysis of RPI prediction tools that do not need experimental HTS data as input. First, we provide a comprehensive overview of available tools grouped by their input requirements. With many RPI prediction tools being developed in recent years, users might lose the overview of accessible tools fitting their needs. To assist potential users, we propose a guide tree covering available RPI prediction algorithms to identify tools of interest for different applications and use cases.

Furthermore, many RPI prediction tools are often benchmarked or usable only with HTS data sets. However, since such data is rarely available for non-model organisms or users might only be interested in a specific RNA-protein complex, we focus on tools that allow such a single-RNA-protein-complex as input. We apply the 30 collected available algorithms on four selected RPI examples across different kingdoms to assess a potential bias towards specific taxonomic clades. Thus, we focus on (i) the human protein LARP7 binding to the 7SK snRNA; (ii) the MS2 phage coat protein interacting with an RNA hairpin in the phage's genome; (iii) the Ebola virus VP30 protein binding to the viral RNA leader region; and (iv) the bacterial toxin-antitoxin system ToxIN. Using this small subset of known RPI examples and their respective “ground truth” of interactions from literature, our evaluation provides a more detailed insight into the capabilities and applicabilities of the RPI prediction algorithms.

2. Results

De novo RPI prediction tools need sequence or structure input of the potentially interacting RNA or protein molecules. Predicted results vary in the degree of information, e.g., some tools only report interaction scores, whereas others report motifs, energies, binding sites, or interacting residues. To assist users in deciding what tool to use for their RPI prediction analysis, we compiled an extensive overview of available RPI prediction tools at GitHub. We further summarized this overview in a guide tree shown in Fig. 1. The tree covers currently (at the time of writing) accessible RPI prediction tools, categorized into necessary input data, computed output, and whether a web server or stand-alone version is available. This overview also contains prediction tools relying on experimental HTS data (see Fig. 1, input data: “experimental dataset”), which were not part of our evaluation. In the following, we review the collected 30 *de novo* RPI prediction tools (see Fig. 1, input data: “RNA”, “protein”, “RNA and protein”) and subsequently present the evaluation results for the available algorithms.

2.1. Overview of de novo RPI prediction tools

RNA sequence input In general, tools using the RNA sequence alone, such as MEME [73], GraphProt [74], and iDeepS [75], report putative sequence motifs of the RNA involved in the interaction. It should be noted that these three tools also have further input options (protein sequence for MEME, HTS datasets for GraphProt and iDeepS). MEME is a motif discovery tool, requiring any kind of biological sequences as input, and calculating (*de novo*) sequence motifs therein, which are displayed and listed for the user. Although not specifically targeted for RPIs, MEME can still serve as an initial step in analyzing potential RNA and protein partners [73]. While both GraphProt and iDeepS work with RNA sequence input, usage still depends on an own HTS dataset to train or the availability of a matching pre-trained model (24 human RBPs for GraphProt, 31 for iDeepS), which was not the case for our chosen biological examples. Therefore, we excluded these two tools from our evaluation. DeepCLIP [76] is a deep neural network trained on CLIP datasets, however, it expects RNA sequence(s) as input, if the user seeks to work with one of the pre-trained models. DeepCLIP calculates RBP binding probabilities for the positions of the given input [76]. RBPmap [64] requires RNA sequence(s) as input as well. Additionally, RBPmap

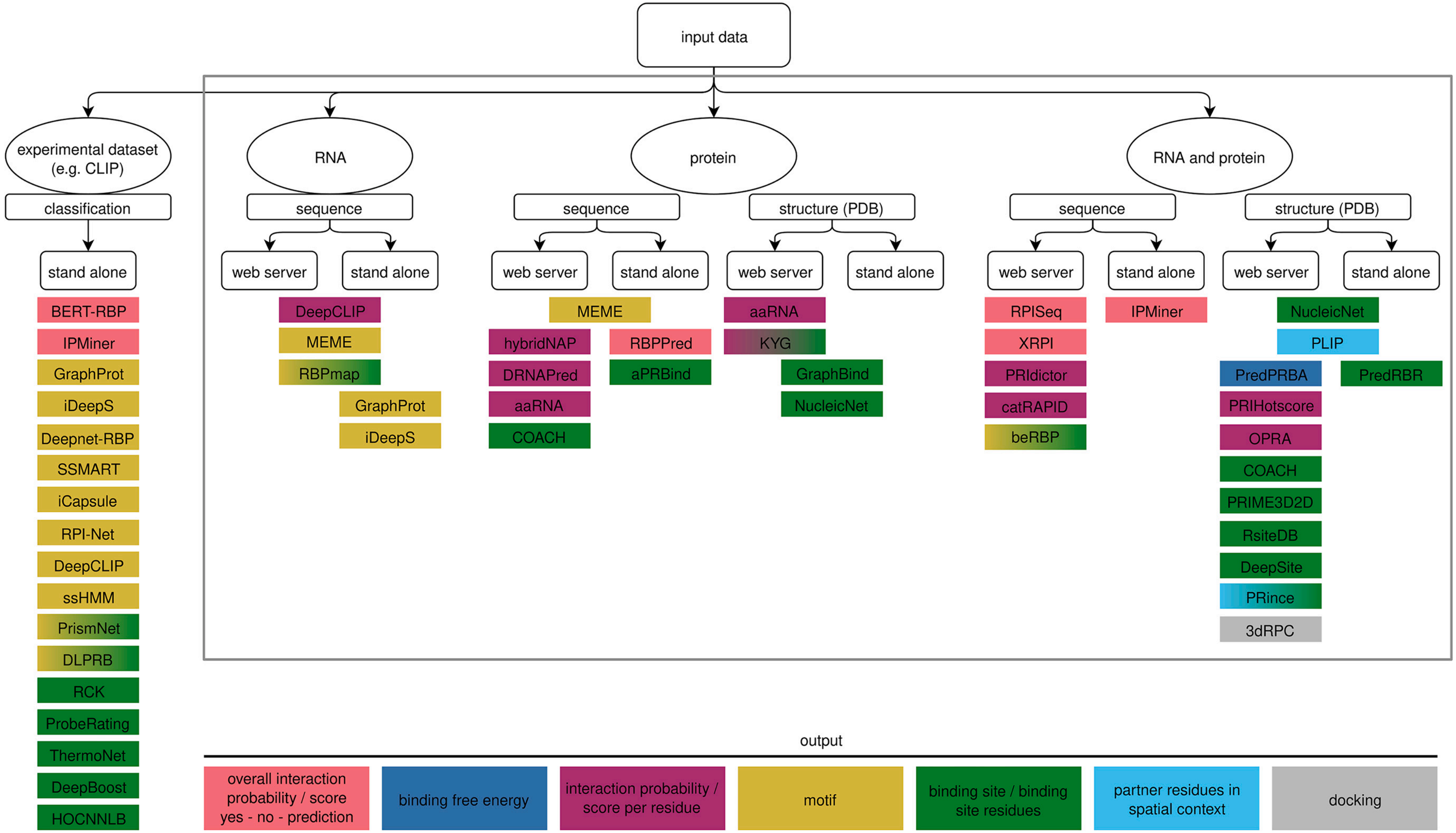

Fig. 1. Guide tree for available RPI prediction tools given specific input data. We provide a (non-exhaustive) list of tools that are still maintained or are at least accessible at the time of writing. Additionally to the input, tools are categorized based on web server or stand-alone versions. The respective output type is color-indicated. A comprehensive tabular overview is also given at GitHub. All tools within the gray-framed box are part of our evaluation in this study. Figure created with InkScape [72].

is able to search for protein-binding motifs in the query RNA(s), which the user has to provide. The motifs are restricted to RBPs in human and two model organisms (*Mus musculus*, *Drosophila melanogaster*).

Protein sequence input If only protein data is available, users can refer to hybridNAP [77], DRNAPred [78], or aaRNA [79]. These tools expect the sequence of the protein and report individual interaction probabilities or scores per amino acid. hybridNAP considers potential binding residues by their relevant sequential properties, relative solvent accessibility in the structure, and their evolutionary conservation, which are assessed based on their test data [77]. These features are covered by DRNAPred as well, in addition to putative intrinsic disorder and secondary structure [78]. The algorithm of aaRNA furthermore includes homology [79]. Similarly, aPRBind [80] annotates residues potentially involved in an interaction using features from protein structure models additionally to the protein sequence, thus trying to incorporate dynamic properties relevant for the potential interaction [80].

Aside from an RNA sequence, MEME can also process protein sequences and reports motifs of interest on the protein. RBPPred [81] allows for the rapid scanning of many proteins by utilizing an SVM classifier to predict whether a protein can bind to an RNA by considering evolutionary information and physicochemical properties of the primary sequence. For each input sequence, the algorithm predicts whether it can bind RNA or not [81].

Protein structure input Given the structure of a protein (most commonly in PDB-like format), GraphBind [82], NucleicNet [67], and KYG [83] are available to predict the binding sites for RNA partners in the protein structure. Users can choose GraphBind to calculate interactions specific for different kinds of ligands. For the prediction, graphs are constructed to reflect the structure context and important features, which are then fed into hierarchical graph neural networks (HGNNs) [82]. NucleicNet predicts and visualizes whether RNA can be bound across the grid of the protein surface given the physicochemical environment, specifically the interaction modes for the different parts of RNA molecules [67]. Furthermore, general binding potential of RNA sequences can be evaluated using logo diagrams. The algorithm is based on the FEATURE vector framework [84], with feature vectors encoding the physicochemical properties. KYG calculates interface residue propensities for each amino acid and residue pairing preferences between the protein and RNA, using data of representative RNA-protein complexes from the PDB [83].

Furthermore, aaRNA can work with a protein structure as well, which results in a structural visualization additionally to the per-residue-probability prediction [79].

RNA and protein sequence input The highest prediction accuracy is possible when RNA and protein information is available. On sequence level, tools such as XRPI [63], RPISeq [62], and IPMiner [85] provide an overall interaction probability, whereas PRIdictor [59] and catRAPID [86] report individual probabilities for residues potentially involved in the interaction. XRPI is a machine-learning method, calculating the interaction probability via a gradient boosting classifier (XGBoost) based on features, such as smallest structural unit and amino acid interaction propensities, from RNA-protein structures in the PDB [63]. The predicted interaction score ranges from 0 to 1. Similarly, RPIseq [62] provides two such scores for a potential RNA and protein sequence pair, calculated by classical machine-learning approaches as well. These two respective classifiers (Support Vector Machine (SVM) and Random Forest (RF)) are trained on non-redundant datasets from the Protein-RNA Interface Database (PRIDB) [87]. The deep learning tool IPMiner extracts features by a stacked autoencoder and uses those for prediction via stacked ensembling of three random forest classifiers [85].

PRIdictor calculates the mutual binding site residues using global and local features of the sequences, as well as partner features, encoded in feature vectors. The predictions consider hydrogen bonds, water bridges, and hydrophobic interactions as potential RPIs [59]. Based on the contributions of hydrogen bonding, van der Waals contacts, and predicted secondary structure of the RNA and protein domains, catRAPID calculates interaction propensities for RNA and protein [86]. With its training data, the algorithm of catRAPID computes and visualizes the pairwise interaction scores for all residues of a given RNA-protein pair and the corresponding discriminative power in a heatmap.

The model of beRBP [88] predicts protein-binding site motifs in given input RNA sequence(s). The user additionally has to provide or choose a pre-trained position weight matrix (PWM) model for an RBP, or an RBP sequence, in which case beRBP tries to predict a corresponding PWM based on similarity of potential RNA binding domains. The 'General model' is trained with a Random Forest approach based on a matrix considering four features of a putative binding site: the match of a motif, sequence environment, spatial accessibility and evolutionary conservation [88].

RNA and protein structure input PredRBR [89], COACH [65], PRIME-3D2D [90], RsiteDB [91,92], and DeepSite [66] use structural information to report binding sites. The accuracy of NucleicNet can be improved (compared to protein structure only) if structural information of both interaction partners is provided.

Although PredRBR [89] works with PDB structures, users depend on the pre-trained model or need to train their own with a respective dataset of structures [89], which is why we excluded this tool from our evaluation. COACH combines multiple algorithms into one approach with a trained SVM classifier to determine consensus binding site residues of a given protein [65]. The tool also allows for a sequence-only input instead of a PDB structure, in which case a 3D model of the protein will be generated prior to the binding site residue prediction. The output lists and visualizes the results of the individual algorithms as well as the combined template-based COACH approach, with the top-ranked models and their calculated confidence score (ranging from 0 to 1) and binding residues [65]. PRIME3D2D refers to structure templates for prediction as well, using TMAlign [93] for protein and LocARNA [94] for RNA alignment to a template to build a RNA-protein complex model, followed by scoring of the potential binding site [90]. RSiteDB [91] focuses on extruded RNA nucleotides not involved in RNA base pairing, which could interact with protein binding pockets. It stores known nucleotide binding sites in RNA-protein structures, as well as provides a prediction service based on the data [92]. The algorithm predicts potential binding sites by determining atomic contacts between the PDB chains, dinucleotide patterns of the RNA, and (geometric) properties of a binding site [92]. DeepSite [66] is a knowledge-based deep convolutional neural network (DCNN) approach that predicts binding sites in proteins for different ligands and includes features such as atom types and chemical properties. The structures are treated like 3D images and ideally cover both the protein and ligands (RNA in our study). The potential residues are calculated and visualized in so-called 'binding site centers' which are supported by a score between 0 and 1.

Our literature search uniquely found PredPRBA [95] to predict the released energy after binding as a measure of possible interaction. The method uses multiple gradient boosted regression tree models for different classes of RNA-protein complexes in the dataset, with features extracted from both sequence and structure [95].

3dRPC [96] is specifically focused on performing docking analyses of RNA and protein molecules into a scored complex. Docking is an alternative computational approach, which was expanded from protein-protein interactions to deal with RPIs only recently [97]. The challenge here lies mostly in the folding flexibility of RNA sequences and the implementation of a scoring function for RNA-protein interactions [97,98].

PRince [99] reports the binding residues of both RNA and protein and produces an output in PDB format that can be used to visualize and explore the structural specificity of the RPI, focused on the accessible surface areas and the highlighted interface region. PRIHotscore [100] and OPRA [101] use the structural information from their respective PDB-derived datasets to report individual interaction probabilities per residue. PRIHotscore approaches the prediction via *in silico* alanine-scanning, which allows to identify interface amino acids as hotspots for RNA binding with assigned interaction scores [100]. OPRA on the other hand assigns interaction propensities based on the ratio of residue composition at RPI interfaces compared to that in the structures' surface using statistical potentials [101].

PLIP [68] finds non-covalent interactions in binding sites between macromolecules or proteins and other biomolecules, like nucleic acids and ligands. Predictions are rule-based on the interaction geometry (i.e., distances and angles) between the residues in a given structure complex, and the physicochemical properties of the amino acids and nucleobases. PLIP reports a detailed list and visualizes the complex with probable interactions, their types, and involved residues and atoms.

2.2. Evaluation of de novo RPI prediction tools

Unfortunately, not all of the 30 described RPI prediction tools could be evaluated to their full extent. As mentioned, GraphProt, iDeepS, and PredRBR depend on pre-trained models or user's HTS datasets (to train a desired model), which is why they were excluded from our evaluation completely (no matching models available). We covered the remaining 27 algorithms, of which multiple ones provided no results to evaluate, due to different individual reasons: (i) not applicable to our examples (no matching data available): RBPmap; (ii) not installable/usable: aPRBind, IPMiner; (iii) computations did not finish or web server did not provide feedback about the status: GraphBind, PRIME3D2D, 3dRPC; (iv) no results are displayed or downloadable: PRince; (v) non-solvable error during input submission: beRBP, RsiteDB.

We introduce an evaluation score for our study, ranging from 0 (no evaluation criterion fulfilled) to 4 (all criteria fulfilled). Since the prediction tools vary widely in their amount and format of output, we decided on four criteria to capture the tool's usefulness for users as best as possible. (I) *Did the algorithm predict any interaction for the given input?*, i.e., depending on the tool, a point is given if the results provide an above-threshold probability of interaction or predict at least one region or residue to interact. (II) *Does the prediction cover the correct region (in protein and/or RNA)?* and (III) *Does the prediction cover the correct interacting residues?*, i.e., if the majority of the literature-based true-positive interacting regions or residues (respectively) are predicted as interacting, a point per criterion is added to the evaluation score. (IV) *Does the tool report a trustable confidence score (>=70%, if applicable)?*, i.e., a point is granted if a confidence score of the tool itself reaches the threshold, or if the amount of true-positive hits outweigh the false-positive and false-negative hits accordingly. During evaluation, these criteria were checked in the mentioned order, with each fulfilled criterion adding 1 point to the score of the respective tool. The nine algorithms with no results to evaluate (listed above) received no assigned evaluation score, leading to 16 tools getting an evaluation score assigned. The full evaluation table and detailed information is deposited at GitHub (`tool_evaluation.xlsx` and `tool_evaluation-score.md`) and includes input, parameters, runtime, output and further information for all tools for our analysis.

2.3. Evaluation dataset: prediction results for four different biological RPIs

The interaction between the LARP7 protein and 7SK RNA is well-described [102] and functions for 7SK stability *in vivo* and assists in a stable association of the 7SK ribonucleoprotein (RNP) complex [110]. The xRRM domain at the C-terminus of LARP7 specifically binds the 3'-terminal U-rich stretch in the 7SK RNA and the top of stem-loop 4 (SL4) [102]. The base G312 and amino acids Y483 and R496 constitute the core interface (Fig. 2A). Additionally, the residues in the three structural elements $\beta 2$ (RNP3 motif), $\beta 3$, and $\alpha 3$ of LARP7 are important for the interaction. The main interaction types are hydrogen bonding and stacking interactions.

The relevant interacting regions for LARP7 and 7SK were predicted most detailed by PLIP and PRIHotscore (evaluation score of 4 and 3, respectively, Fig. 3). Both managed to positively predict the RPI involved residues while correctly excluding non-interacting ones. Since the available PDB complex represents the specific RPI region well, other tools considering structure also performed well with scores of 2: COACH, NucleicNet, and OPRA, as well as MEME using sequence only. The most important residues are predicted by multiple tools, but rarely the RPI is covered exhaustively or in full detail. For the protein residues, we observed multiple false positive predictions, i.e., amino acids predicted as interacting although not explicitly being mentioned as such in the corresponding literature [102]. This can be seen for example with aaRNA, hybridNAP, KYG, and catRAPID. For the latter, interactions are focused at the 5'-end of the 7SK RNA and across the complete LARP7 sequence, with an overall high interaction propensity of 84. As LARP7 binds to the UUU-3'OH and the SL4 stem-loop [111], the prediction at the 5'-end is not supported by the literature. Moreover, *in vivo*, the 5'-end of 7SK RNA is bound by the methylphosphate capping enzyme (MePCE) [102], denying the opportunity for interaction with LARP7 in this region. Another crucial aspect in prediction for structure-based tools is the presence of the RNA molecule in the 3D complex. Exemplarily, since the LARP7 PDB entry does not include the 7SK RNA structure, the results of DeepSite are less reliable than if the RNA was present, i.e., the prediction of true positive amino acids for interaction is restricted (here: only the helix and C-terminal end of $\alpha 3$ is predicted correctly). In the case of a model-based approach like COACH, the prediction is highly dependent on the availability of adequate models for the example at hand. The top-ranked model for LARP7 is based on a Polypyrimidine tract-binding protein PTB (PDB:2ADC, covers an RRM domain) with a CUCUCU-RNA strand as a ligand. Since the interacting stem-loop of 7SK consists of the motif AUGAUG, the U-richness might be reflected in the model but only allows limited conclusions to the interaction of LARP7 with 7SK RNA. PRIdictor only predicted a few interacting residues (1 amino acid, 10 nucleobases) with its web application and none with the web server, therefore getting a score of 0.5.

Another well-studied RPI example is the coat protein of the RNA phage MS2. As a dimer, it interacts with a 19-nucleotide-long RNA region in the phage genome, which folds into a hairpin structure and contains the initiation code for the replicase gene (Fig. 2B). The interaction of RNA and protein leads to a switch from replication to virion assembly in the viral life cycle [103,112].

The interaction was detected most reliably and correctly by COACH, hybridNAP, PRIHotscore, and PLIP (Fig. 3). All of them correctly predict the majority of the RPI-involved amino acids [104] in the protein with high probability. However, they also predict false positive residues, which cannot be distinguished without knowing the 'ground truth' (hybridNAP, COACH), or miss important interaction partners (PRIHotscore, PLIP). With a score of 2, aaRNA, RPISeq, and XRPI respectively perform better for this RPI compared to the other biological examples. aaRNA predicts binary binding propensities above threshold for some but not all interacting amino acids, and both RPISeq and XRPI confidently deliver high interaction probabilities with both classifiers (SVM: 0.96, RF: 1.0) or models (RPI2825: 0.9969, RPI390: 0.9191), respectively.

Our third RPI is the N-terminal region of VP30 in Ebola virus binding to a stem-loop structure in the 3'-leader region of the single-stranded, negative-oriented RNA genome at nucleotides 80-54 and the complementary antigenomic strand (Fig. 2C). The RPI is focused on an arginine-rich region in the protein and regulates the transcription of the virus [106,107].

The best ranking tools here, with a score of 2, were COACH, OPRA, and PRIdictor (Fig. 3). The latter does predict a subset of important interacting residues, but no regions as a whole, and also covers multiple false positive hits. COACH and OPRA mostly do not provide RPI residues, which is in line with the input structure not including the main interacting region and was therefore the expected outcome.

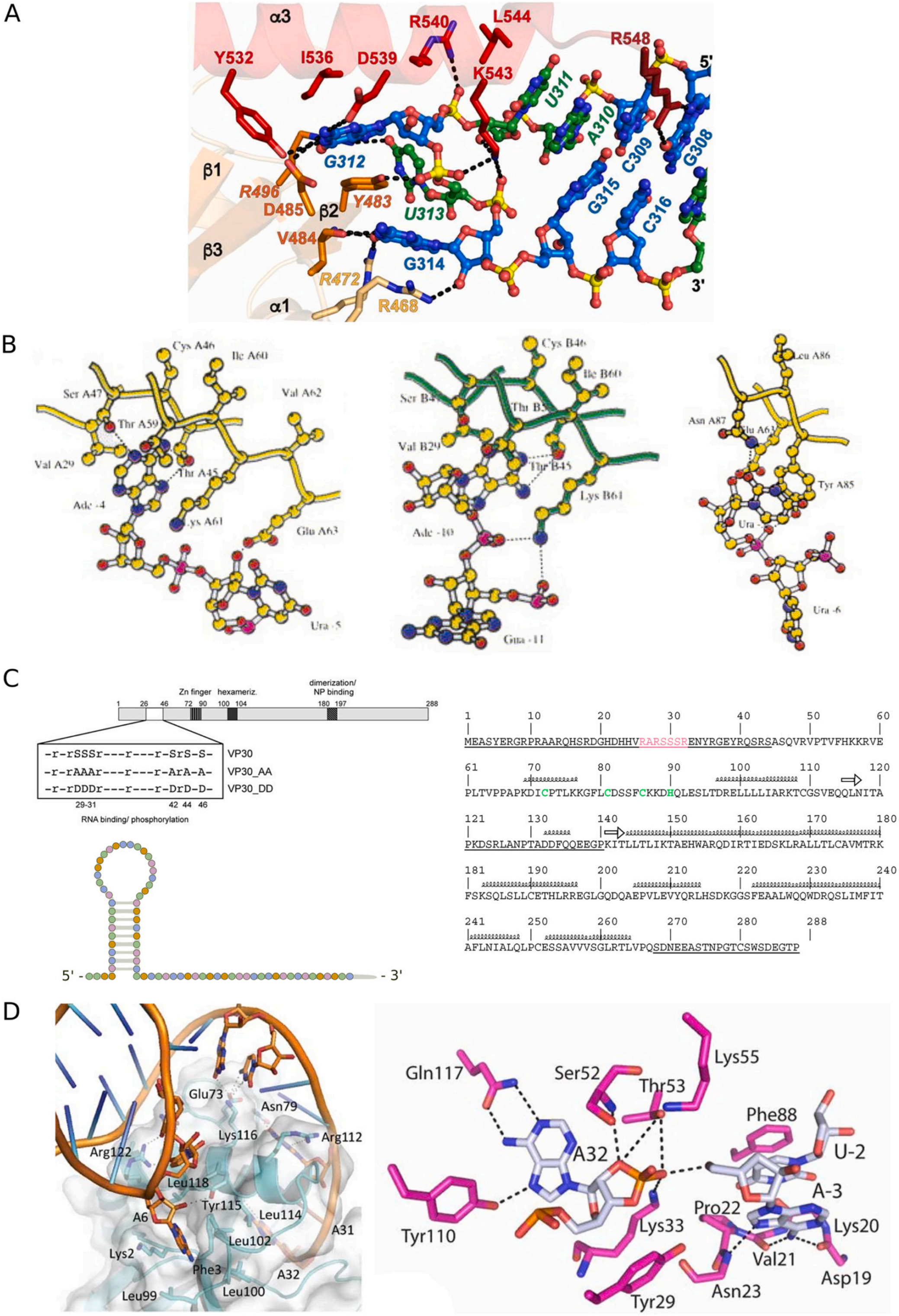

Fig. 2. Known interactions for the four investigated RPI examples based on the respective literature. (A) The C-terminus of LARP7 (xRRM domain) binds the RNA with key amino acids in three structural elements ($\beta 2$, $\beta 3$, $\alpha 3$). The 7SK RNA is bound at the stem-loop 4 (SL4) and a U-rich region at the 3'-end. Interacting residues are shown as a stick representation, colored by structure (protein) or element (RNA). Figure from Eichhorn et al. (2018) [102]. **(B)** A 19-nucleotide-long RNA region in the MS2 phage genome folds into a hairpin structure bound at exposed positions by a dimer of the MS2 phage coat protein. The contacting amino acids (different binding pockets depicted here) are primarily conserved within the β-strands of the protein structure [103]. Figures from Valegård et al. (1997) [104]. **(C)** The N-terminal region of the Ebola virus VP30 protein binds the 3'-leader region of the RNA genome strand at nucleotides 80-54 with an arginine-rich region (highlighted in white) [105] (top left). The optimal RNA substrate is single-stranded, of 40nt length and mixed base composition [106] (bottom left). Further single amino acids contribute to the RPI indicated in pink and green [107] right. Figures from Biedenkopf et al. (2016), Schlereth et al. (2016) (adapted), John et al. (2007) [105–107]. **(D)** The RPIs in the bacterial ToxIN system assist in assembling the heterohexameric complex of three toxI RNA pseudoknots and three ToxN monomers. The interactions are described to be focused around a few key nucleobases bound by various amino acids of the protein in multiple pockets. Shown are a hydrophobic binding pocket of ToxN (left) and the active site of the complex. Figures from Blower et al. (2011), Short et al. (2012) [108,109]. For full details, please refer to the respectively cited original publications.

The algorithms of catRAPID, hybridNAP, MEME, RBPPred, RPISeq, and XRPI all predict some kind of interaction or motif (depending on their output type) based on the input sequence(s), however it does not represent the literature-known residues [106,107] (score 1).

Multiple structure-based tools cannot predict the RPI due to the disadvantageous PDB complex. This leads to the algorithms either reporting interactions despite the PDB entry not covering the interaction-important N-terminal region of the protein (score of 0: KYG, aaRNA, NucleicNet, DeepSite) or not being able to start their computations at all because of the lacking RNA molecule in the structure (no score: PLIP, PredPRBA, PRIHotscore).

As a fourth example, we investigated the bacterial toxin-antitoxin system ToxIN, which encodes an ABI (abortive infection) system in multiple, especially enteric bacteria [113]. To neutralize the toxin protein (ToxN), three ToxI RNA pseudoknots form a heterohexameric complex with three ToxN monomers mediated by RPIs in different binding pockets [108]. The interactions are focused around multiple amino acids, which group together to bind a few nucleobases, respectively (Fig. 2D).

Here, DeepSite, and again PRIHotscore and PLIP perform the best with scores of 3 and 4 (Fig. 3). PLIP correctly identifies nearly all interaction-involved residues for both RNA and protein and their respective relations. DeepSite and PRIHotscore perform similarly, but are missing a few relevant amino acids. Both COACH and OPRA provide multiple false positive or false negative matches, respectively, in addition to some of the literature-known amino acids [108]. The two classifiers of RPISeq as well as the two models of XRPI come to different interaction probabilities each, which led us to assign half a point for both tools. RBPPred did predict the ToxN sequence as non-interacting, as opposed to its results for the three other examples.

3. Discussion

3.1. *Quality (and quantity) of predictions differs greatly between different tools*

The investigated RPI prediction tools work with different algorithms and on different data. Thus, their results vary vastly in their informative value and level of detail. Working with sequence input, the results of RBPPred, XRPI, and RPISeq provide a general single score or probability with no information on specific residues or regions. This might prove useful in case the user just needs this distinction of how likely an interaction is, e.g., for an “all-against-all” approach to compare multiple RNA and protein sequences. For RPISeq no obvious relation between the two classifiers is stated, complicating the interpretation whether an RPI can be trusted in the case of two diverging predicted probabilities. In such instances, the user cannot know if the RNA and protein are likely to interact, especially if there is no ground truth available. The motif finder MEME overall covers interacting regions in our evaluation, however the motifs are distributed across the sequences in general, challenging the motif’s specificity. Although MEME does not predict RPIs directly, it might prove useful if a motif of interest is known before analysis or to narrow down a potential amount of possible interaction sites for further experiments [73]. For probabilities or scores of individual residues in the sequence users can refer to DeepCLIP, hybridNAP, DRNApred, PRIdictor, or catRAPID. The latter additionally correlates every residue of both the RNA and protein sequence against each other. DRNApred does not predict any interactions for any of our examples and should therefore be used with caution.

Since spatial factors play crucial roles in RPIs [24,25], structure-based tools generally produce more detailed results. The binding free energy calculated by PredPRBA provides an overall tendency for the RPI complex of interest. For per-residue-information, users can apply KYG, aaRNA, PRIHotscore, or OPRA, all of which supplement this with visualization in the structural complex. Specific binding sites are reported with NucleicNet, DeepSite, and COACH. The former highlights binding

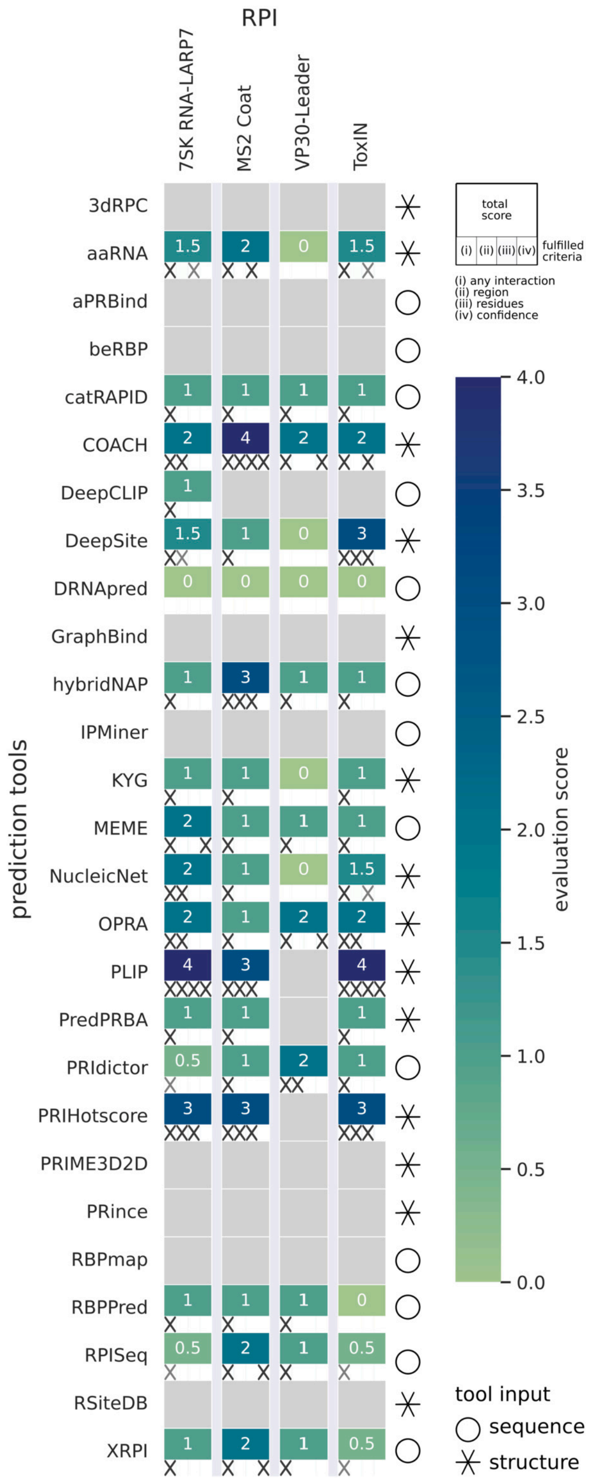

Fig. 3. Evaluation heatmap. Based on the criteria listed in the Methods, we evaluated the RPI prediction tools for all four biological examples. The score therefore can range from 0 (lowest/worst) to 4 (highest/best) (color scale), or not assigned (gray) because of the reasons listed in section “Evaluation of *de novo* RPI prediction tools”. For each tool and example combination, the cross notation indicates which of the criteria are fulfilled respectively. Gray crosses indicate 0.5 points. Tools marked with a circle require sequence input and the star indicates a structure input. For an overview of installation and running accessibility, we kindly refer to Table `tool_evaluation.xlsx` (GitHub repository).

areas of potential interaction, whereas the latter two further indicate the respective amino acids involved. The multiple result models provided by COACH give the user the option to consider predictions with (potentially) low confidence scores, which may still prove useful for discovering “hidden” RPI-relevant amino acids. Without a ground truth, the manual evaluation of multiple models might be prone to errors. PLIP is the only one with information on both RNA and protein residues in its output, and also shows the types of interactions and therefore grants the most detailed predictions for all examples.

For all tools, data availability is essential. While sequence data is far more abundant than structural data for RPI complexes, the latter is mandatory for structure-dependent tools. A general problem in structure determination is the flexibility of the RPI complex [22], which is typically captured in one state only and thus might not reflect all aspects of the interaction. Working with datasets of bound and unbound states of the proteins could support RPI prediction in this aspect. Additionally, only the protein structure of an RPI complex without the RNA partner is commonly determined, which impedes the structure-based algorithms to predict the RPI accurately. Furthermore, as seen in the Ebola virus VP30 example, similar issues arise when the protein structure is incomplete, especially when domains or regions important for the interactions are missing. This underlines the need for more structural data from experimental determination methods and continuous improvement of computational tools to predict as much accurate information as possible with as little data as necessary.

Here, we focused on *de novo* prediction tools instead of tools using or analyzing experimental RPI datasets (e.g. CLIP-based). The latter usually requires an additional training before the actual prediction to fit the model to the organism of interest. Especially for non-bioinformaticians, building and training a model on robust data sets is a challenging task and thus hinders the usage of such tools. Accessibility and usability are key factors in providing interpretable *in silico* RPI predictions.

3.2. No tool is more accurate across all kingdoms

We examined the RPI prediction tools with RPI examples from different species: LARP7 and 7SK RNA acting in humans, MS2 phage coat protein and an RNA hairpin in the MS2 phage genome, VP30 protein and viral RNA leader region of Ebola virus, and the bacterial toxin-antitoxin system ToxIN. We did not observe drastic accuracy differences between the investigated tools in one of the examples. Some algorithms seemed to perform better (according to the evaluation score) for the MS2 phage coat protein RPI, namely hybridNAP, aaRNA, XRPI, and COACH, while others were challenged by the VP30 RPI, as seen for KYG, aaRNA, NucleicNet, PredPRBA, and DeepSite. Problems with the structure-based prediction for VP30 most likely stem from the PDB entries covering only parts of the protein and no RNA at all. Furthermore, the structure-based tools PLIP, PredPRBA, and PRIHotscore depend on the RNA structure in the PDB complex for their predictions. Including all structural components of the RPI generally grants more sound predictions, but is hindered if there is no adequate data available for the user to apply. Overall, in this investigation, prediction accuracies depended more on the functionality and input of tools (sequence versus structure) than the origin of biological examples. However, due to the restricted availability of training data, some tools, such as the proposed framework by Shulman-Peleg et al. [91], might be biased towards human or model organism RPIs or RPIs that are easier to investigate with the standard experimental methods. Consequently, these tools might perform more reliably for such RPIs, but might not generalize on non-model organisms and underrepresented RPIs. Even predictions within the organisms used for training models may fail to generalize for RPI instances, as with catRAPID (trained on human RPIs) and the human LARP7 and 7SK RNA example. RBPmap, RsiteDB, and beRBP are trained on human data as well, but since we could not evaluate results from those tools, we cannot deduce whether they would have been biased in favor of the human example. Similarly, algorithms providing pre-trained models for prediction are based on the available data and therefore focused on human data as well. For example, in our evaluation DeepCLIP only performs for the 7SK RNA, since LARP7 is the only protein available as a pre-trained model. Furthermore, the restricted data availability extends also to the input for the tools, as mentioned above. The higher quality and amount of information the input data has, the more accurate the prediction tools can calculate potential RPIs, regardless of the species or kingdom.

3.3. Post-processing potential of RPI prediction tools

Besides the prediction results themselves, the tools differ regarding how well their output can be used for further downstream analyses. Only displaying the results at the web server without the possibility for download of important output data hinders the user from meaningfully integrating a prediction tool into a self-build pipeline. We ranked NucleicNet, PRIdictor, catRAPID, and DeepSite as tools with a low post-processing potential. The annotated PyMol [114] session file provided by NucleicNet and the heatmap plot by catRAPID are downloadable, but only represent the predictions visually without the potential to parse the data. Besides the plot visualizations, PRIdictor only provides a text file for the protein predictions, not the RNA partner, which does not represent the result data in a well-retrievable way. DeepSite lets the user download the spatial coordinates of the ‘interaction centers’, not the interacting residues themselves.

Many tools provide their RPI prediction results in one or multiple parsable file formats, in addition to visual representations, depending on the respective tool. This includes DeepCLIP, MEME, hybridNAP, DRNAPred, aaRNA, COACH, KYG, PLIP, PRIHotscore, and OPRA. RBPPred, RPISeq, XRPI, and PredPRBA result in just a score value, which is downloadable as a text file for the latter two.

We cannot make statements regarding the post-processing potential for RBPmap, aPRBind, GraphBind, IPMiner, beRBP, PRIME3D2D, RsiteDB, PRince, and 3dRPC, as these tools either did not complete their calculations, did not process the input, or it was not possible to get them running (see Results).

4. Conclusion

RPIs are ubiquitous in all life forms and can be studied with experimental detection methods and bioinformatic prediction algorithms based on their interaction features. With the growing amount of available data, current models and approaches can be improved on or expanded with this data to support RPI research and understanding further. In recent years, many tools and workflows have been proposed to predict RPIs with and without the requirement of HTS data. While this is a positive development for the field, users might lose the overview of accessible tools fitting their needs.

In this study, we provide an overview of RPI prediction tools and proposed a guide tree. We structured this overview and tree according to the tools’ required input and the degree of detail produced by their output. With this evaluation, we provide a guide for users to support the identification of appropriate tools for their research.

To assess the reliability of tools utilizing machine- and deep-learning techniques, we investigated the algorithms in more detail using four known RPI examples covering different kingdoms of life. The tools report varying amounts of detail and information about the specifics of the respective interaction. By having a ground truth at hand, this study gives insights into the reliability and interpretability of the tools. Computational predictions always have to be evaluated carefully, and this study is not exhaustive in terms of possible RPI examples and the performance of tested tools. Without prior knowledge, not all tools are valuable for *de novo* prediction use cases. Low confidence of the algorithms complicates interpretation of results or separation of false positives and false negatives from true predictions. The structure-based tools overall provided more details on the interactions, but rely on the availability of RNA-protein structure complexes.

Table 1
Overview of RPI examples and respective input for the tools. The table lists the RPI examples used in this study and the literature describing known interacting residues. Furthermore, it shows which data (sequence/structure) was used for the prediction tool evaluation.

Examples	Organism	RPI ref.	Input sequence	Input structure
7SK RNA	human	[102]	NCBI Gene ID 125050 (NC_000006.12: 52995620-52995951)	PDB 6D12 PDB 5KNW (DeepSite only)
LARP7			UniProt Q4G0J3	
MS2 phage operator RNA	MS2 phage	[104]	PDB 1ZDH sequence GenBank MK213795.1, nt 1730-1779 (catRAPID only)	PDB 1ZDH
MS2 phage coat protein			UniProt P03612	
Ebola genomic RNA	Ebola virus	[106,107]	Schlereth et al. [106], nt 153-4	PDB 5DVW
VP30			UniProt Q05323	
ToxI (RNA)	*Pectobacterium atrosepticum*	[108,109]	PDB 2XD0 sequence (repeated to be 50nt long for catRAPID only)	PDB 2XD0 PDB 4ATO (KYG, NucleicNet)
ToxN (protein)			UniProt B8X8Z0	

Due to limited RPI data, predictions might be biased toward the interaction mechanisms of the organisms used for training and benchmarking. Many available tools rely on training data originating from human or model organism data. To overcome these issues, homology-based approaches, including evolutionary information, as intended in COACH or aaRNA, may be suited. However, additional and continuous refinement of such models is needed when new data is available. Thus, to deepen our understanding of RPIs and their exact mechanisms, we expect to see a continuing close exchange between experimental detection and *in silico* prediction models in the future, i.e., experimental data being fed into existing or new algorithms, while *in silico* prediction results reduce the space of potential RPI interactions to investigate in the lab and thus the necessary time and costs.

5. Materials and methods

5.1. RPI data selection

We used four cross-species examples for the RPI tool evaluation: (i) human protein LARP7 with the 7SK snRNA, (ii) MS2 phage coat protein with an RNA hairpin in the phage's genome, (iii) Ebola virus VP30 protein with the viral RNA leader region, and (iv) bacterial toxin-antitoxin system ToxIN. The examples are based on how well they are researched, whether there is available data to use for the tools, and to cover species from different kingdoms.

For algorithms requiring a 3D structural complex, we provided a corresponding PDB entry structure for each example (see Table 1). For the 7SK RNA-LARP7-complex, we chose the PDB:6D12 structure entry because it specifically contains the interacting regions of RNA and protein. However, DeepSite could not resolve the necessary features of this structure (error for protonation of amino acids), which is why we used PDB:5KNW (same as PDB:6D12 but without RNA) for this tool instead. Both structures only represent the relevant RNA-interacting xRRM domain of LARP7 and not the whole protein structure. Determined structures for VP30, unfortunately, currently only cover the protein partially (5DVW starts with position 139 compared to the UniProt sequence), and lack RNA molecules. The preferred PDB entry for the ToxIN complex was 2XD0 which comprises the full heterohexameric molecule, but for KYG and NucleicNet, we had to choose PDB:4ATO (only one protein and RNA monomer each) due to complications of these tools with the former structure.

For the sequence-based tools, we downloaded the protein sequences from the respective UniProt entries (see Table 1). The sequence of the 7SK RNA corresponds to NCBI Gene ID:125050. For the MS2 operator RNA, we used the sequence from PDB:1ZDH (the wild-type U/T at position 11 has been substituted with C for improved interaction and crystallization process [104]). Since catRAPID needs at least 50nt for input, we used the corresponding GenBank entry (MK213795.1) to add 16 upstream and 15 downstream nucleotides to the sequence. We retrieved the interacting genomic, negative-oriented RNA strand of Ebola virus from the supplementary of Schlereth et al., covering nucleotides 153-4 [106], which contains the most important positions for the RPI. We picked the RNA sequence of ToxI from entry PDB:2XD0. We extended this entry for catRAPID by appending the first 14 nucleotides of the sequence to the 3'-end to reach 50nt total. We justify this with the biological function: The RNA sequence is present as multiple repeats in the bacterium before being cut into the interacting monomers [113].

5.2. Evaluation of tool results

The details regarding web server or stand alone usage, as well as version number, parameters, runtime, input, output, are listed in Table `tool_evaluation.xlsx` (GitHub repository). We evaluated the results with the support of information about the respective RPIs described in the literature, gained in experiments. The knowledge originates from introducing point substitutions in connection with the determination of the equilibrium dissociation constant (K_d) using ITC (isothermal titration calorimetry) [102], crystallization experiments and structure determination [102,104,108,109], deletion mutants and site-directed mutations [106,107], or EMSA analysis [106].

We decided on the following evaluation criteria:

- Did the algorithm predict any interaction for the given input?
- Does the prediction cover the correct region (in protein and/or RNA)?
- Does the prediction cover the correct interacting residues?
- Does the tool report a trustable confidence score (>=70%, if applicable)?

Because of the great differences of format and amount of output provided by the algorithms, selection of these criteria proved difficult. Fulfillment of each criterion leads to 4 points total maximum. A score of zero implies that the tool did provide prediction results, but none of the evaluation criteria were fulfilled. No score implies that the tool was not applicable to the respective example, did not complete its computations, or could not be installed. Table `tool_evaluation.xlsx` (GitHub repository) gives an overview of the tool evaluation and relevant information, and detailed evaluation score distribution for each tool and example is noted in Table `tool_evaluation-score.md` (GitHub repository). The evaluation heatmap was plotted with an in-house python script and finalized with InkScape v.1.1.2 [72].

Furthermore, we assessed the potential for post-processing the results of each tool, e.g., for downstream analysis after the RPI prediction. (+) denotes parsable, downloadable results the user can potentially incorporate into a computational pipeline. Tools with a (-) provide viewable content on their web server, but no option for download or potential to incorporate the results directly into downstream analyses of the user. (n/a) marks the tools which were not applicable (no evaluation score) and we thus could not assess those regarding their post-processing potential.

CRediT authorship contribution statement

Sarah Krautwurst: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Kevin Lamkiewicz:** Writing – review & editing, Validation, Project administration, Investigation, Conceptualization.

Supporting information

Supporting information can be found at GitHub.

Declaration of competing interest

The authors have declared no conflict of interest.

Acknowledgements

We thank Dr. Emanuel Barth and Prof. Dr. Manja Marz for proofreading the manuscript.

References

[1] Re A, Joshi T, Kulberkyte E, et al. RNA–protein interactions: an overview. In: Gorodkin J, Ruzzo WL, editors. RNA sequence, structure, and function: computational and bioinformatic methods. Totowa, NJ: Humana Press; 2014. p. 491–521.

[2] Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet 2014;15(12):829–45.

[3] Holmqvist E, Vogel J. RNA-binding proteins in bacteria. Nat Rev Microbiol 2018;16(10):601–15.

[4] Babitzke P, Baker CS, Romeo T. Regulation of translation initiation by RNA binding proteins. Annu Rev Microbiol 2009;63(1):27–44.

[5] Pullmann R, Kim HH, Abdelmohsen K, et al. Analysis of turnover and translation regulatory RNA-binding protein expression through binding to cognate mRNAs. Mol Cell Biol 2007;27(18):6265–78.

[6] Dassi E. Handshakes and fights: the regulatory interplay of RNA-binding proteins. Front Mol Biosci 2017;4.

[7] Zealy RW, Wrenn SP, Davila S, et al. microRNA–binding proteins: specificity and function. WIREs RNA 2017;8(5).

[8] Jiang P, Coller H. Functional interactions between microRNAs and RNA binding proteins. MicroRNA 2012;1(1):70–9.

[9] Li J-H, Liu S, Zheng L-L, et al. Discovery of protein-lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets. Front Bioeng Biotechnol 2015;2.

[10] Noh JH, Kim KM, McClusky WG, et al. Cytoplasmic functions of long noncoding RNAs. WIREs RNA 2018;9(3).

[11] Lukong KE, Chang K-w, Khandjian EW, et al. RNA-binding proteins in human genetic disease. Trends Genet 2008;24(8):416–25.

[12] Musunuru K. Cell-specific RNA-binding proteins in human disease. Trends Cardiovasc Med 2003;13(5):188–95.

[13] Zhou H, Mangelsdorf M, Liu J, et al. RNA-binding proteins in neurological diseases. Sci China Life Sci 2014;57(4):432–44.

[14] Helder S, Blythe AJ, Bond CS, et al. Determinants of affinity and specificity in RNA-binding proteins. Curr Opin Struct Biol 2016;38:83–91.

[15] Hentze MW, Castello A, Schwarzl T, et al. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol 2018;19(5):327–41.

[16] Smirnov A, Förstner KU, Holmqvist E, et al. Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. Proc Natl Acad Sci USA 2016;113(41):11591–6.

[17] Girardi E, Pfeffer S, Baumert TF, et al. Roadblocks and fast tracks: how RNA binding proteins affect the viral RNA journey in the cell. Semin Cell Dev Biol 2021;111:86–100.

[18] Williamson JR. Induced fit in RNA-protein recognition. Nat Struct Biol 2000;7(10):834–7.

[19] Leulliot N, Varani G. Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. Biochemistry 2001;40(27):7947–56.

[20] Hainzl T, Huang S, Sauer-Eriksson AE. Structural insights into SRP RNA: an induced fit mechanism for SRP assembly. RNA 2005;11(7):1043–50.

[21] Zhang T, Zhang H, Chen K, et al. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. Curr Protein Pept Sci 2010;11(7):609–28.

[22] Gunasekaran K, Nussinov R. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. J Mol Biol 2007;365(1):257–73.

[23] Liu N, Dai Q, Zheng G, et al. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature 2015;518(7540):560–4.

[24] Jones S, Daley DT, Luscombe NM, et al. Protein-RNA interactions: a structural analysis. Nucleic Acids Res 2001;29(4):943–54.

[25] Liu S, Li B, Liang Q, et al. Classification and function of RNA-protein interactions. Wiley Interdiscip Rev RNA 2020;11(6):e1601.

[26] Vuković L, Koh HR, Myong S, et al. Substrate recognition and specificity of double-stranded RNA binding proteins. Biochemistry 2014;53(21):3457–66.

[27] Jankowsky E, Harris ME. Specificity and nonspecificity in RNA-protein interactions. Nat Rev Mol Cell Biol 2015;16(9):533–44.

[28] Corley M, Burns MC, Yeo GW. How RNA-binding proteins interact with RNA: molecules and mechanisms. Mol Cell 2020;78(1):9–29.

[29] Ellis JJ, Broom M, Jones S. Protein-RNA interactions: structural analysis and functional classes. Proteins 2007;66(4):903–11.

[30] Marchese D, de Groot NS, Lorenzo Gotor N, et al. Advances in the characterization of RNA-binding proteins. Wiley Interdiscip Rev RNA 2016;7(6):793–810.

[31] Ramanathan M, Porter DF, Khavari PA. Methods to study RNA-protein interactions. Nat Methods 2019;16(3):225–34.

[32] Hogg JR, Collins K. RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. RNA 2007;13(6):868–80.

[33] Tsai BP, Wang X, Huang L, et al. Quantitative profiling of in vivo-assembled RNA-protein complexes using a novel integrated proteomic approach. Mol Cell Proteomics 2011;10(4):M110.007385.

[34] Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc 2006;1(1):302–7.

[35] Ule J, Jensen K, Mele A, et al. CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods 2005;37(4):376–86.

[36] Dominguez D, Freese P, Alexis MS, et al. Sequence, structure, and context preferences of human RNA binding proteins. Mol Cell 2018;70(5):854–8679.e.

[37] Taliaferro JM, Lambert NJ, Sudmant PH, et al. RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol Cell 2016;64(2):294–306.

[38] Lambert N, Robertson A, Jangi M, et al. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol Cell 2014;54(5):887–900.

[39] Mili S, Steitz JA. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. RNA 2004;10(11):1692–4.

[40] Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotechnol 2011;29(7):607–14.

[41] Han Y, Guo X, Zhang T, et al. Development of an RNA-protein crosslinker to capture protein interactions with diverse RNA structures in cells. RNA 2021;28(3):390–9.

[42] Zhao J, Ohsumi TK, Kung JT, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell 2010;40(6):939–53.

[43] Licatalosi DD, Mele A, Fak JJ, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 2008;456(7221):464–9.

[44] Ikehata H, Ono T. The mechanisms of UV mutagenesis. J Radiat Res 2011;52(2):115–25.

[45] Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdiscip Rev RNA 2010;1(2):266–86.

[46] Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010;141(1):129–41.

[47] Huppertz I, Attig J, D'Ambrogio A, et al. iCLIP: protein-RNA interactions at nucleotide resolution. Methods 2014;65(3):274–87.

[48] Hafner M, Landthaler M, Burger L, et al. PAR-CliP – a method to identify transcriptome-wide the binding sites of RNA binding proteins. J Vis Exp 2010;41.

[49] Konig J, Zarnack K, Rot G, et al. iCLIP–transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. J Vis Exp 2011;50.

[50] Blencowe BJ, Ahmad S, Lee LJ. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. Genes Dev 2009;23(12):1379–86.

[51] Derrien T, Estelle J, Marco Sola S, et al. Fast computation and applications of genome mappability. PLoS ONE 2012;7(1):e30377.

[52] Corcoran DL, Georgiev S, Mukherjee N, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol 2011;12(8):R79.

[53] Uren PJ, Bahrami-Samani E, Burns SC, et al. Site identification in high-throughput RNA-protein interaction data. Bioinformatics 2012;28(23):3013–20.

[54] Guo JK, Blanco MR, Walkup 4th WG, et al. Denaturing purifications demonstrate that PRC2 and other widely reported chromatin proteins do not appear to bind directly to RNA in vivo. Mol Cell 2024;84(7):1271–128912.e.

[55] Sagar A, Xue B. Recent advances in machine learning based prediction of RNA-protein interactions. Prot Peptide Lett 2019;26(8):601–19.

[56] Munteanu A, Mukherjee N, Ohler U. SSMART: sequence-structure motif identification for RNA-binding proteins. Bioinformatics 2018;34(23):3990–8.
[57] Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res 2020;30(2):214–26.
[58] Singh H, Raghava GPS. BLAST-based structural annotation of protein residues using Protein Data Bank. Biol Direct 2016;11(1):4.
[59] Tuvshinjargal N, Lee W, Park B, et al. PRIdictor: protein-RNA interaction predictor. Biosystems 2016;139:17–22.
[60] Zhang W, Qu Q, Zhang Y, et al. The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. Neurocomputing 2018;273:526–34.
[61] Berman HM, Battistuz T, Bhat TN, et al. The protein data bank. Acta Crystallogr, Sect D, Biol Crystallogr 2002;58(6):899–907.
[62] Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. BMC Bioinform 2011;12:489.
[63] Jain DS, Gupte SR, Aduri R. A data driven model for predicting RNA-protein interactions based on gradient boosting machine. Sci Rep 2018;8(1):9552.
[64] Paz I, Kosti I, Ares Jr M, et al. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res 2014;42(Web Server issue):W361–7.
[65] Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 2013;29(20):2588–95.
[66] Jiménez J, Doerr S, Martínez-Rosell G, et al. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics 2017;33(19):3036–42.
[67] Lam JH, Li Y, Zhu L, et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. Nat Commun 2019;10(1):4941.
[68] Adasme MF, Linnemann KL, Bolz SN, et al. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Res 2021;49(W1):W530–4.
[69] Si J, Cui J, Cheng J, et al. Computational prediction of RNA-binding proteins and binding sites. Int J Mol Sci 2015;16(11):26303–17.
[70] Yan J, Zhu M. A review about RNA–protein-binding sites prediction based on deep learning. IEEE Access 2020;8:150929–44.
[71] Wei J, Chen S, Zong L, et al. Protein-RNA interaction prediction with deep learning: structure matters. Brief Bioinform 2022;23(1).
[72] Inkscape: open source scalable vector graphics editor; 2021.
[73] Bailey TL, Johnson J, Grant CE, et al. The MEME suite. Nucleic Acids Res 2015;43(W1):W39–49.
[74] Maticzka D, Lange SJ, Costa F, et al. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol 2014;15(1):R17.
[75] Pan X, Rijnbeek P, Yan J, et al. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics 2018;19(1):511.
[76] Grønning AGB, Doktor TK, Larsen SJ, et al. DeepCLIP: predicting the effect of mutations on protein-RNA binding with deep learning. Nucleic Acids Res 2020;48(13):7099–118.
[77] Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. Brief Bioinform 2017;20(4):1250–68.
[78] Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 2017;45(10):e84.
[79] Li S, Yamashita K, Amada KM, et al. Quantifying sequence and structural features of protein-RNA interactions. Nucleic Acids Res 2014;42(15):10086–98.
[80] Liu Y, Gong W, Zhao Y, et al. aPRBind: protein-RNA interface prediction by combining sequence and I-TASSER model-based structural features learned with convolutional neural networks. Bioinformatics 2021;37(7):937–42.
[81] Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM. Bioinformatics 2017;33(6):854–62.
[82] Xia Y, Xia C-Q, Pan X, et al. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. Nucleic Acids Res 2021;49(9):e51.
[83] Kim OTP, 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(22):6450–60.
[84] Halperin I, Glazer DS, Wu S, et al. The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. BMC Genomics 2008;9(S2).
[85] Pan X, Fan Y-X, Yan J, et al. IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genomics 2016;17:582.
[86] Agostini F, Zanzoni A, Klus P, et al. catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. Bioinformatics 2013;29(22):2928–30.
[87] Lewis BA, Walia RR, Terribilini M, et al. PRIDB: a protein-RNA interface database. Nucleic Acids Res 2011;39(Database issue):D277–82.
[88] Yu H, Wang J, Sheng Q, et al. beRBP: binding estimation for human RNA-binding proteins. Nucleic Acids Res 2018;47(5):e26.
[89] Tang Y, Liu D, Wang Z, et al. A boosting approach for prediction of protein-RNA binding residues. BMC Bioinform 2017;18(Suppl 13):465.
[90] Xie J, Zheng J, Hong X, et al. PRIME-3D2D is a 3D2D model to predict binding sites of protein-RNA interaction. Commun Biol 2020;3(1):384.
[91] Shulman-Peleg A, Shatsky M, Nussinov R, et al. Prediction of interacting single-stranded RNA bases by protein-binding patterns. J Mol Biol 2008;379(2):299–316.
[92] Shulman-Peleg A, Nussinov R, Wolfson HJ. RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. Nucleic Acids Res 2009;37(Database):D369–73.
[93] Zhang Y. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–9.
[94] Will S, Reiche K, Hofacker IL, et al. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol 2007;3(4):e65.
[95] Deng L, Yang W, Liu H. PredPRBA: prediction of protein-RNA binding affinity using gradient boosted regression trees. Front Genet 2019;10:637.
[96] Huang Y, Liu S, Guo D, et al. A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Sci Rep 2013;3:1887.
[97] Nithin C, Ghosh P, Bujnicki JM. Bioinformatics tools and benchmarks for computational docking and 3D structure prediction of RNA-protein complexes. Genes 2018;9(9).
[98] Puton T, Kozlowski L, Tuszynska I, et al. Computational methods for prediction of protein–RNA interactions. J Struct Biol 2012;179(3):261–8.
[99] Barik A, Mishra A, Bahadur RP. PRince: a web server for structural and physicochemical analysis of protein-RNA interface. Nucleic Acids Res 2012;40(Web Server issue):W440–4.
[100] Krüger DM, Neubacher S, Grossmann TN. Protein-RNA interactions: structural characteristics and hotspot amino acids. RNA 2018;24(11):1457–65.
[101] 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(1):25–35.
[102] Eichhorn CD, Yang Y, Repeta L, et al. Structural basis for recognition of human 7SK long noncoding RNA by the La-related protein Larp7. Proc Natl Acad Sci USA 2018;115(28):E6457–66.
[103] Valegård K, Murray JB, Stockley PG, et al. Crystal structure of an RNA bacteriophage coat protein-operator complex. Nature 1994;371(6498):623–6.
[104] Valegård K, Murray JB, Stonehouse NJ, et al. The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. J Mol Biol 1997;270(5):724–38.
[105] Biedenkopf N, Schlereth J, Grünweller A, et al. RNA binding of Ebola virus VP30 is essential for activating viral transcription. J Virol 2016;90(16):7481–96.
[106] Schlereth J, Grünweller A, Biedenkopf N, et al. RNA binding specificity of Ebola virus transcription factor VP30. RNA Biol 2016;13(9):783–98.
[107] John SP, Wang T, Steffen S, et al. Ebola virus VP30 is an RNA binding protein. J Virol 2007;81(17):8967–76.
[108] Blower TR, Pei XY, Short FL, et al. A processed noncoding RNA regulates an altruistic bacterial antiviral system. Nat Struct Mol Biol 2011;18(2):185–90.
[109] Short FL, Pei XY, Blower TR, et al. Selectivity and self-assembly in the control of a bacterial toxin by an antitoxic noncoding RNA pseudoknot. Proc Natl Acad Sci USA 2013;110(3):E241–9.
[110] Markert A, Grimm M, Martinez J, et al. The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. EMBO Rep 2008;9(6):569–75.
[111] Muniz L, Egloff S, Kiss T. RNA elements directing in vivo assembly of the 7SK/MePCE/Larp7 transcriptional regulatory snRNP. Nucleic Acids Res 2013;41(8):4686–98.
[112] Rūmnieks J, Tārs K. Protein-RNA interactions in the single-stranded RNA bacteriophages. In: Harris JR, Bhella D, editors. Virus protein and nucleoprotein complexes. Singapore: Springer Singapore; 2018. p. 281–303.
[113] Fineran PC, Blower TR, Foulds IJ, et al. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. Proc Natl Acad Sci USA 2009;106(3):894–9.
[114] Schrödinger LLC. The PyMOL molecular graphics system, version 2.0; 2015.