PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

RNANetMotif: Identifying sequence-structure RNA network motifs in RNA-protein binding sites

Hongli Ma[1,2,3,4☯], Han Wen[2☯], Zhiyuan Xue[5], Guojun Li[1,4,6*], Zhaolei Zhang[2,3,7*]

1 Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China, **2** Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada, **3** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, **4** School of Mathematics, Shandong University, Jinan, China, **5** West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China, **6** School of Mathematical Science, Liaocheng University, Liaocheng, China, **7** Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

☯ These authors contributed equally to this work.
* guojunsdu@gmail.com (GL); Zhaolei.Zhang@utoronto.ca (ZZ)

OPEN ACCESS

Citation: Ma H, Wen H, Xue Z, Li G, Zhang Z (2022) RNANetMotif: Identifying sequence-structure RNA network motifs in RNA-protein binding sites. PLoS Comput Biol 18(7): e1010293. https://doi.org/10.1371/journal.pcbi.1010293

Editor: Arjun Krishnan, Michigan State University, UNITED STATES

Received: September 16, 2021

Accepted: June 9, 2022

Published: July 12, 2022

Data Availability Statement: The RNANetMotif software is available for download at https://github.com/hongli-ma/RNANetMotif. eCLIP datasets used in this study can be found at "Download" page on our webserver (http://rnanetmotif.ccbr.utoronto.ca).

Funding: GL acknowledges support from National Science Foundation of China (NSFC, 11931008; 61771009) and The National Key Research and Development Program of China (No. 2020YFA0712400). ZZ acknowledges funding from Natural Sciences and Engineering Research

Abstract

RNA molecules can adopt stable secondary and tertiary structures, which are essential in mediating physical interactions with other partners such as RNA binding proteins (RBPs) and in carrying out their cellular functions. In vivo and in vitro experiments such as RNAcompete and eCLIP have revealed in vitro binding preferences of RBPs to RNA oligomers and in vivo binding sites in cells. Analysis of these binding data showed that the structure properties of the RNAs in these binding sites are important determinants of the binding events; however, it has been a challenge to incorporate the structure information into an interpretable model. Here we describe a new approach, RNANetMotif, which takes predicted secondary structure of thousands of RNA sequences bound by an RBP as input and uses a graph theory approach to recognize enriched subgraphs. These enriched subgraphs are in essence shared sequence-structure elements that are important in RBP-RNA binding. To validate our approach, we performed RNA structure modeling via coarse-grained molecular dynamics folding simulations for selected 4 RBPs, and RNA-protein docking for LIN28B. The simulation results, e.g., solvent accessibility and energetics, further support the biological relevance of the discovered network subgraphs.

Author summary

RNA binding proteins (RBPs) regulate every aspect of RNA biology, including splicing, translation, transportation, and degradation. High-throughput technologies such as eCLIP have identified thousands of binding sites for a given RBP throughout the genome. It has been shown by earlier studies that, in addition to nucleotide sequences, the structure and conformation of RNAs also play important role in RBP-RNA interactions. Analogous to protein-protein interactions or protein-DNA interactions, it is likely that there exist intrinsic sequence-structure motifs common to these RNAs that underlie their binding

Council of Canada (Discovery Grant, RGPIN-2017-06743). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: none

specificity to specific RBPs. It is known that RNAs form energetically favorable secondary structures, which can be represented as graphs, with nucleotides being nodes and backbone covalent bonds and base-pairing hydrogen bonds representing edges. We hypothesize that these graphs can be mined by graph theory approaches to identify sequence-structure motifs as enriched sub-graphs. In this article, we described the details of this approach, termed RNANetMotif and associated new concepts, namely EKS (Extended K-mer Subgraph) and GraphK graph algorithm. To test the utility of our approach, we conducted 3D structure modeling of selected RNA sequences through molecular dynamics (MD) folding simulation and evaluated the significance of the discovered RNA motifs by comparing their spatial exposure with other regions on the RNA. We believe that this approach has the novelty of treating the RNA sequence as a graph and RBP binding sites as enriched subgraph, which has broader applications beyond RBP-RNA interactions.

1. Introduction

The human genome encodes approximately 1500 or more RNA binding proteins (RBPs), which regulate every aspect of the RNA biogenesis and RNA biology, including RNA splicing, modification, degradation, protein translation, and RNA subcellular localization [1–5]. RBPs are also involved in many important developmental processes such as embryogenesis, proliferation, and differentiation. Mutations or dysregulation of RBPs are also implicated in many human diseases including cancer [6–8]. In vitro methods such as RNAcompete, RNA Bind-n-Seq (RBNS) and high-throughput RNA-SELEX (HTR-SELEX) can measure binding affinity of an RBP to RNA oligomers [9–12], and in vivo methods such as PAR-CLIP, iCLIP, and eCLIP (enhanced UV crosslinking followed by immunoprecipitation) can identify regions on the RNA transcript that are bound by RBPs in cells [1,13–15]. Analysis of these in vitro binding affinity data and in vivo binding sites have revealed that the structure of RNA molecules can help understand the binding mode between a specific RBP and their target sequences. For example, RNA binding domains can be grouped into single stranded or double stranded RNA binding domains (ssRBD and dsRBD) based on their preference for RNA targets that are either single stranded (unpaired) or double stranded (paired). A number of computational methods have been developed to ascertain the structure properties of these RBP bound RNAs with the aim to incorporate structural information into a predictive model that can help decode mechanisms that regulate protein-RNA interactions [16–35]. A number of these representative methods are briefly described below.

These structure-based methods differ in how they encode RNA structure information in their models thus they can be grouped into two broad categories according to the abstract level of structural encoding. Methods like MEMERIS [18], mCarts [29] and SMARTIV [22] apply simplified structure representations, including RNA accessibility scores and discrete secondary structure notations such as paired and unpaired, while methods including RNAcontext [17], ssHMM [24], BEAM [25], GraphProt [16] and SARNAclust [27] characterize the shape of substructures with different labeling methods. Approaches such as RNAcontext [17] explicitly label individual nucleotides with an additional feature indicating the secondary structure of the particular nucleotide, i.e., paired, hairpin loop. Methods such as GraphProt [16] and ssHMM [24] further capture structure inter-dependencies between neighboring nucleotides through graph based encoding or hidden Markov model. Moreover, methods such as BEAM [25] attempts to explicitly encode the pattern of nucleotide base pairing and catalogues the occurrence of specific secondary structure motifs based on the length of the stem or the size of the hairpin loop. SARNAclust [27] comprehensively studies the topology and annotations of the complete RNA structure and provides several options for graph transformations of RNA shapes. A more recent method, RPI-Net, uses graph neural network to represent RBP bound RNAs and uses a deep learning approach to distinguish bound and non-bound RNAs [34]. Despite these advances, our understanding on the structural basis of binding events between protein and RNA still lags our understanding on the interaction between protein partners, which is largely due to the lack of atomic resolution RNA structures.

In parallel to the study of protein-RNA interactions, our understanding on protein-DNA interactions has been greatly aided by the availability of a plethora of structure information on protein-DNA complexes and on DNA structures. It has been recognized that the overall shapes of the DNA molecules play an important role in the initial recognition by proteins and the subsequent binding processes [36,37]. These structural elements allow optimal positioning between amino acids and the interacting moieties on DNA molecules, e.g., backbone or nucleosides. It is likely that similar principles are also important in protein-RNA recognition, i.e., there exist recurring RNA 2D or 3D structure motifs that are essential in the stabilization of the RNA molecule and important in presenting the nucleotide moieties to RBPs [38–41]. The 2D structure of an RNA molecule consists of a network of base-pairing interactions between nucleotides, which, without atomic resolution structure, is the best approximation of a RNA's 3D structure. We hypothesize that there exist intrinsic subgraphs in these networks that can potentially separate RNA sequences that are bound by a specific RBP or RBPs from those that are not bound; we term these 2D sequence entities as RNA network motifs.

Interactions between RNA-binding proteins (RBPs) and RNA molecules employ diverse and dynamic modes [38,42]; finding the exact structural motifs has been a challenging problem. There have been previous attempts trying to solve this problem in an approximate way [16,19,24,25]. Despite the recent advances, there remain two issues that need to be addressed. First, we need to combine sequence and structural information of RNA in an efficient and flexible way. Global approaches such as GraphProt [16] model RNA structure as complete graphs; other methods such as BEAM [25] use an additional alphabet to represent the structural state of each nucleotide. A framework that can both capture the states of individual nucleotides while encoding the base-pairing information between the nucleotides would be more desired. Secondly, given that RNA structures are represented as 2D or 3D graphs, we need efficient and robust approaches to search many of these input structures and identify enriched and potentially discriminative network motifs.

Towards these goals, we herein introduce a novel algorithm, RNANetMotif, which takes as input thousands of RNA sequences presumably bound by an RBP and uses a graph theory approach to search for "base pair derived subnetworks" enriched in the predicted RNA secondary structures. RNANetMotif consists of the following steps. (i) For each RNA locus bound by an RBP as determined via eCLIP experiments, we predict its base-pairing pattern using software RNAplfold and represent its secondary structure as a network [43]. (ii) We developed a novel GraphK algorithm, which can partition the aforementioned RNA secondary structure network into subgraphs consisted of both RNA backbone and base pair interactions. For each RBP, these subgraphs are pooled and filtered to obtain a candidate pool. (iii) We compute HVDM (Heterogeneous Value Difference Metric) distance matrix to construct a similarity network with candidate subgraphs being nodes, and further conducted network pruning. (iv) In the similarity network, we next perform maximal clique enumeration and rank the occurrence of nodes in large maximal cliques to obtain enriched subgraph candidates, i.e., RNA net motifs. (v) Finally, we conducted 3D structure modeling of selected RNA sequences through coarse-grained molecular dynamics (MD) folding simulation and evaluated the significance of the discovered network motifs by comparing their spatial exposure with other regions [44].

We note, in addition to RBP binding sites, RNANetMotif can also be applied in other scenarios to extract enriched RNA sequence or structure motifs. RNANetMotif compares favorably with other methods in terms of performance and run-time. We believe RNANetMotif offers a fresh approach in understanding interactions between RBP and RNA targets, and in extracting informative RNA motifs. To make this tool more accessible to the community, we constructed a web server (http://rnanetmotif.ccbr.utoronto.ca) that stores the results from our analysis of ENCODE RBP datasets (16 RBPs) and allows users to upload a collection of RNA sequences of their own interests for motif analysis. The RNANetMotif is a new way of investigating RNA sequences and can be further extended into analysis of other categories of RNA sequences.

2. Results

2.1 Overall workflow of the RNANetMotif method

As shown in **Fig 1A and 1B**, RNANetMotif consists of the following steps. In Step 1, we predict RNA base-pairing probability using RNAplfold and represent each RNA binding site as a graph in which edges represent base pairings and backbones. In Step 2, we developed a GraphK algorithm to partition the network into Extended K-mer Subgraphs (EKSes) and to filter to obtain the final EKS pool. In Step 3, we compute the HVDM (Heterogeneous Value Difference Metric) distance matrix [45] and use network pruning to construct a similarity network among the EKSes. In Step 4, we identify overlapping densely connected modules on the similarity network and evaluate the significance of these modules by comparing them with a previously established negative set. We next identify the representative EKSes by conducting maximal clique enumeration, followed by extracting the overlapping large size cliques in the similarity network. In Step 5, for several RBPs, we conducted 3D structure modeling on the RNAs to further examine the network motifs in a structure and dynamic context. As an example, we selected LIN28B protein and performed protein-RNA docking to examine the detailed binding mode and evaluated the robustness of the RNANetMotif method.

2.2 Selection and optimization of *k* in GraphK partition

As shown in **Fig 1A** **(Step 2)** and **Fig 2**, the EKSes were constructed by traversing and extending *k* nucleotides from a specific base-pair (see **Methods** **4.3, 4.4** for details). To select the most appropriate *k* to partition the RNA molecule, we investigated the occurrence of intervals between predicted base-pairs in the predicted RNA secondary structure.

We first divide all the nucleotides in the entire RNA as paired or unpaired and define the distance between two adjacent paired nucleotides as gap length. As shown in **Fig 3** (top panel), approximately 79.51% of all the gap lengths is 1, which represents stacked adjacent base pairs. The bottom of **Fig 3** shows, excluding gap length 1, gap length < = 9 can cover 87.91% of all possible gaps in 22 RBPs. Note that EKSes were constructed by extending from the initial base pair in two directions, thus a gap length of 9 is equivalent to 5-mer extension. This indicates that, with *k* ranging from 3 to 5, 87.91% of the nucleotides in RBP binding sites are likely covered by at least one EKS. We also note that most RNA binding domains tend to bind to short RNA sequences, therefore we investigated *k* up to 5 in our pipeline. The RNANetMotif pipeline is quite flexible so users can customize the cut-off *k* or other global parameters to suit their requirements. Moreover, the bottom bar plot in **Fig 3** shows that the number of occurrences decreases as the gap length increases.

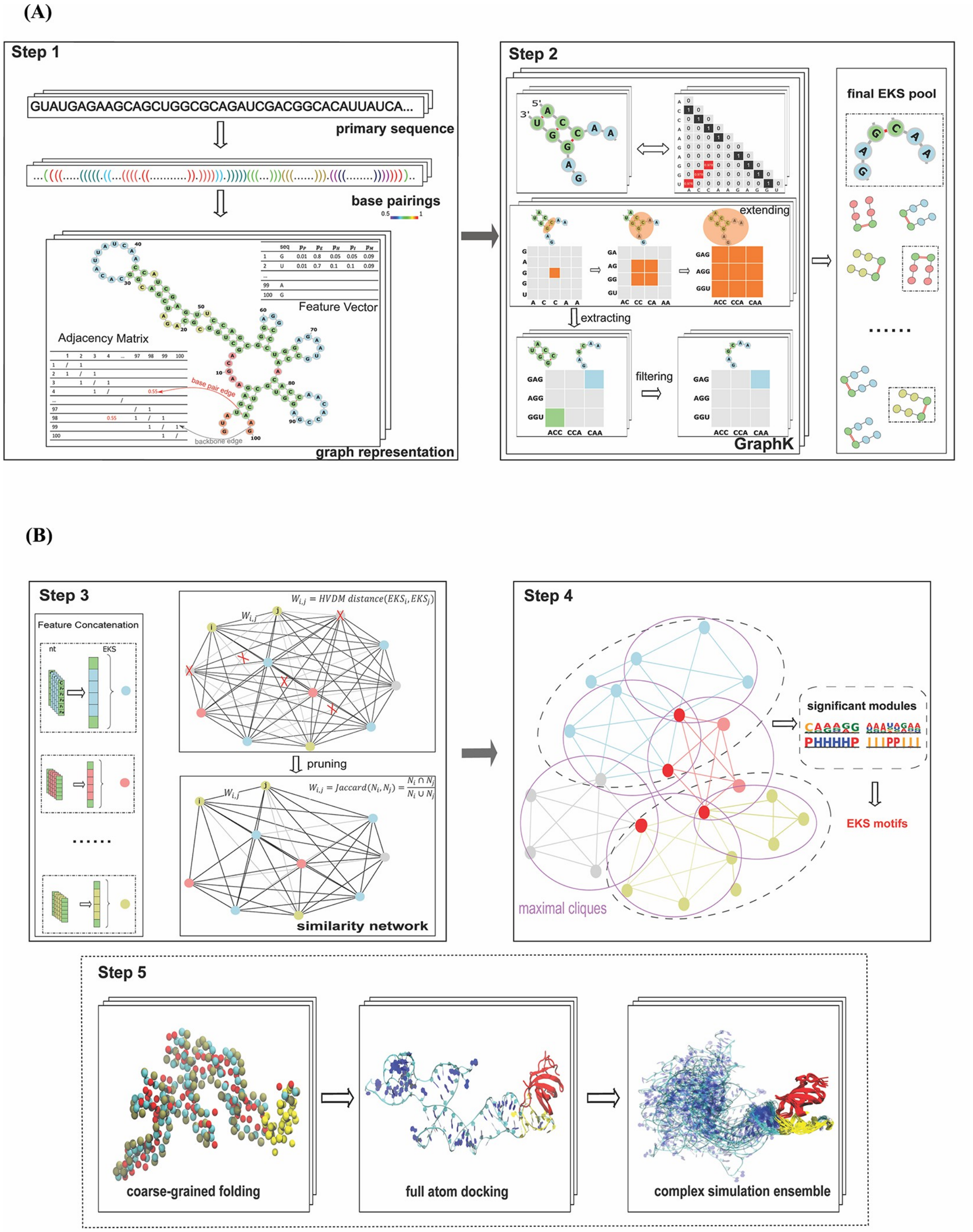

Fig 1. **(A)** Workflow of RNANetMotif (First Part). **Step1.** Predicting base-pairings of protein-bound RNA sequences and graph representation. The representation includes nucleotide information (vertex feature vector) and base-pairing information (adjacency matrix), here depicted with different colors to distinguish backbone links (grey) from base pairing (red). **Step 2.** GraphK partition algorithm (see Methods) to obtain final EKS pool. **(B).** Workflow of RNANetMotif (Second Part). **Step3.** Calculate HVDM (Heterogeneous Value Difference Metric) distance matrix and construct similarity network of EKSes. **Step 4.** Detect significant network modules and then identify intrinsic EKS motifs. **Step5.** RNA 3D structure modeling via discrete molecular dynamics folding simulations and protein-RNA docking with simulation for validation.

https://doi.org/10.1371/journal.pcbi.1010293.g001

After applying GraphK to partition the RBP binding sites into EKSes, we pooled these selected EKSes and investigated the properties of these elements. We further discretized the predicted secondary structure probability of each nucleotide into one of the following 5 states:

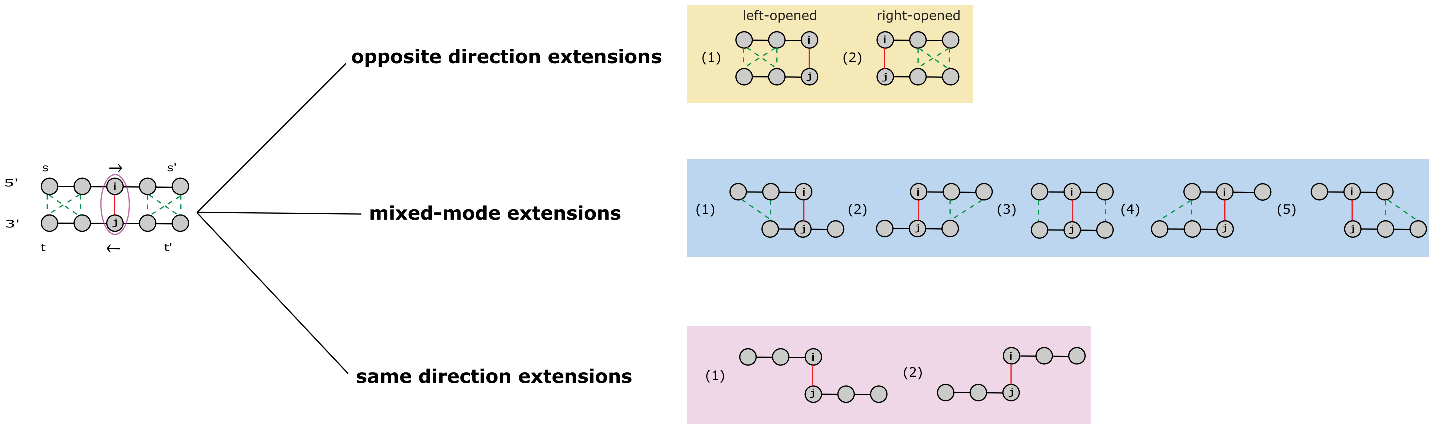

Fig 2. Definition and classification of EKSes. As displayed, there are three categories of EKS: opposite-direction extensions (right-opened and left-opened), mixed-mode extensions and same-direction extensions. Black edges represent backbone bonds, red edges represent base pair interaction between i and j, green dashed edges represent possible base pair interactions.

https://doi.org/10.1371/journal.pcbi.1010293.g002

P (paired), H (hairpin loop), E (extended, unstructured), I (internal loop, bulge), and M (multiple loop). We next counted the occurrence of sequence and structure of EKSes respectively; Fig 4 shows the frequency of top sequence instances and top secondary structure instances in scatter plots for five selected RBPs on different *k*. From Fig 4, we find that after applying GraphK partition, the distribution of top sequence instances of EKS has lower frequencies

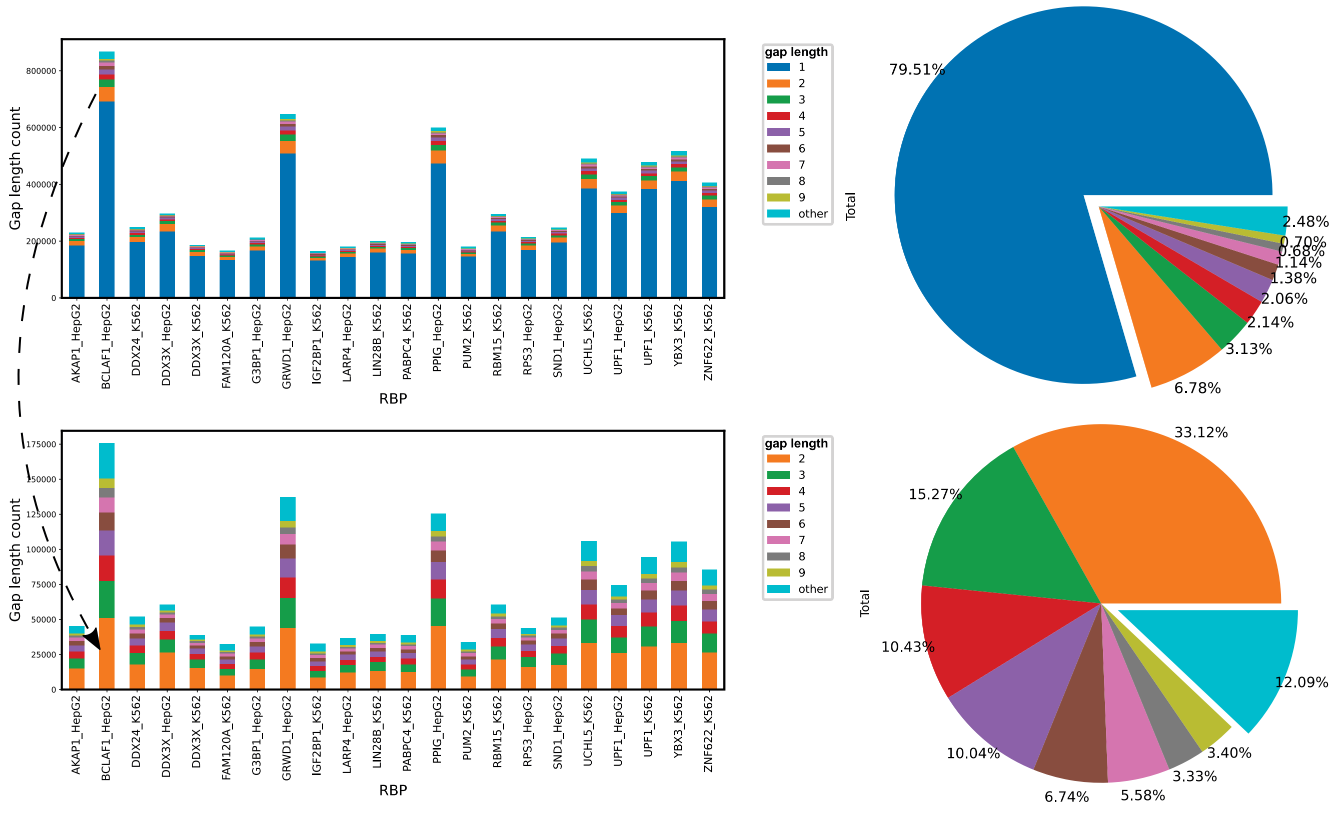

Fig 3. Distribution of gap lengths among base-pairings for individual RBPs.

https://doi.org/10.1371/journal.pcbi.1010293.g003

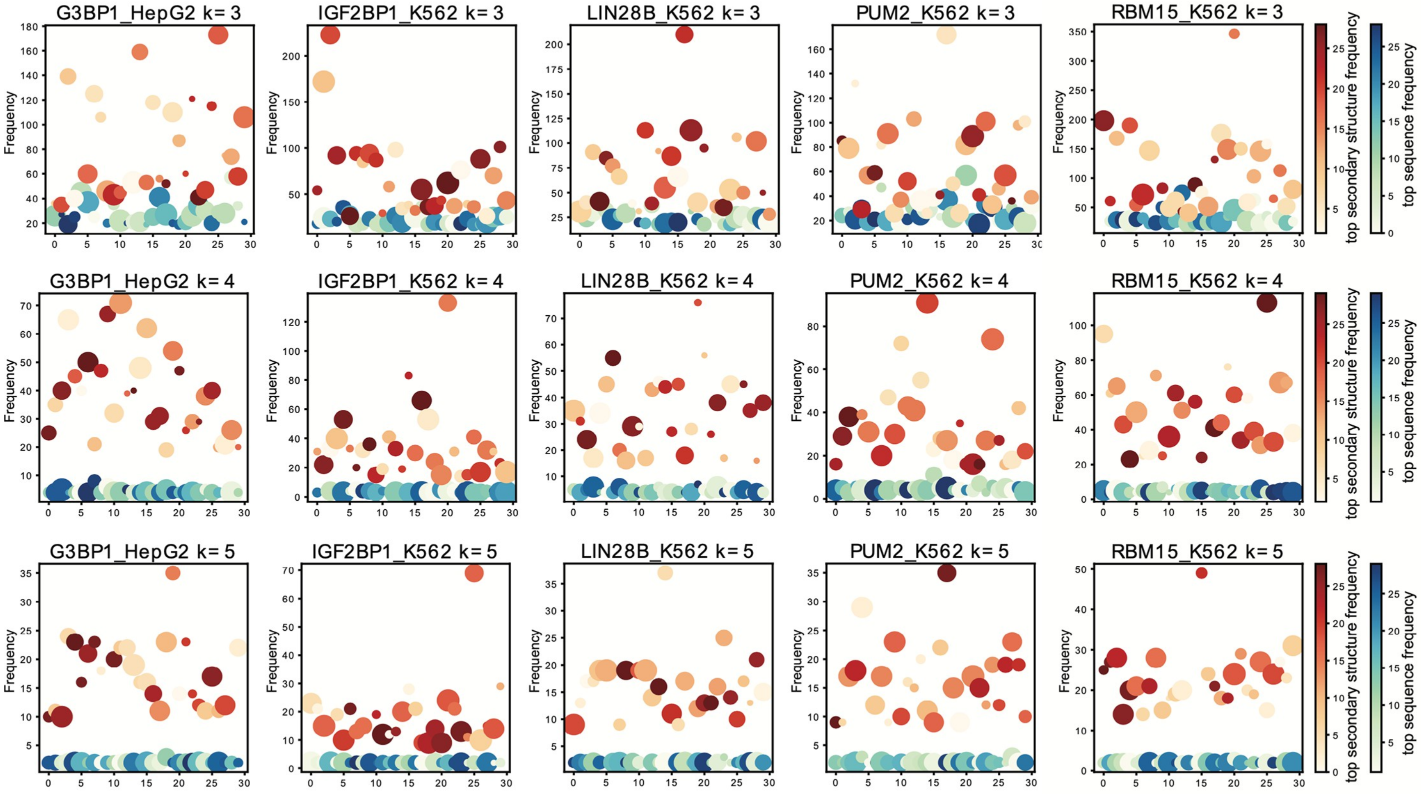

Fig 4. Distribution of the frequency of top sequence- and structure- instances in final EKS pool for 5 RBPs.

https://doi.org/10.1371/journal.pcbi.1010293.g004

than top structure instances. Given the low number of top sequence instances occurrence, we fitted a new distance and enrichment definition in our graph-preserving framework (see **Methods 4.5**, **4.6**).

2.3 Significant modules in the similarity network of EKSes

We next calculated HVDM distances among pairs of EKSes and constructed a similarity network with individual EKSes as nodes and HVDM distances as edge weights. We then used ClusterOne software to search for densely connected network modules in the similarity network and evaluate the statistical significance of these modules by Pearson χ^2 test [46]. As described above in **Section 2.2**, we next discretized secondary structure representation of each nucleotide and visualized the sequence-structure preference of these significant modules in **Fig 5** for 16 selected RBPs. We did not observe statistically significant modules for RBPs that preferentially bind to double-stranded RNAs such as DDX3X, DDX24, GRWD1 (containing WD repeat domain). We think this may be because RNA structure plays different roles in protein-RNA recognition for double-stranded and single-stranded RBPs, which is consistent with previous observations [47, 48]. We also did not observe significant EKS modules for two other RBPs, PPIG and BCLAF1, which may be because these RBPs have multiple binding modes or multiple RNA binding domains, generating a mixture of signals [49]. Further investigation is warranted for these RBPs.

Fig 5 shows the sequence logo and associated secondary structures of the enriched EKS modules for different k-mer sizes and EKS topologies. We notice that proteins with the same

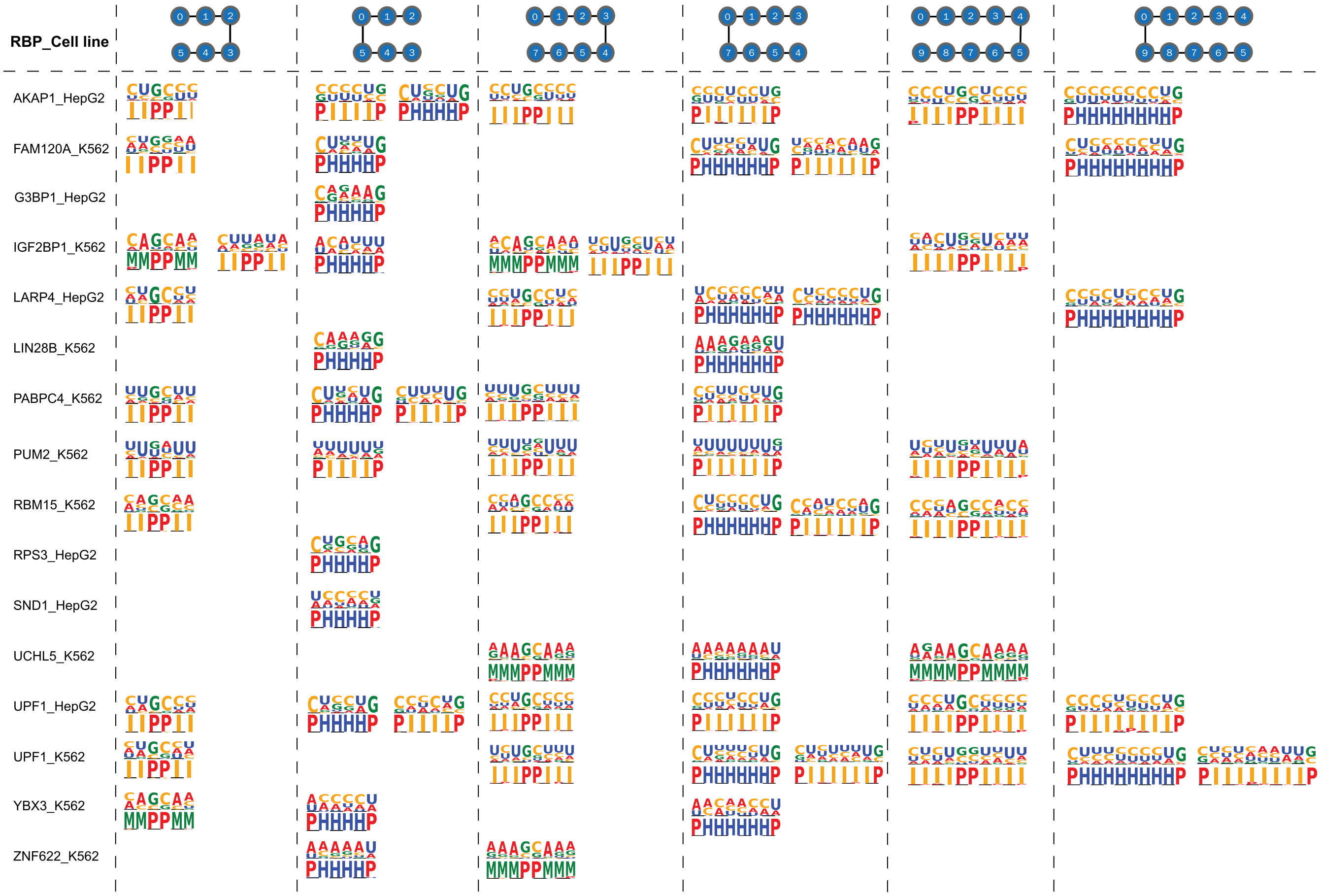

Fig 5. Combined sequence and structure logos of significant modules of 16 RBPs. Left-opened and right-opened EKSes of different sizes are displayed.

https://doi.org/10.1371/journal.pcbi.1010293.g005

binding domains tend to have similar binding profiles. For example, IGF2BP1, PABPC4, LARP4 and RBM15 share a common RRM (RNA recognition Motif) domain, consisted of pyrimidine-rich internal loops and hairpin loops. Not surprisingly, we also note that the same proteins have similar profiles in different cell lines. For instance, profiles of UPF1 from HepG2 and K562 tend to share similar sequence preferences in both hairpin loops and internal loops when *k* is set at 5.

2.4 Significance of EKS motifs tested by 3D structure modelling

In the world of proteins, linear polypeptides usually fold into certain three-dimensional structures to ultimately carry out cellular functions. Similarly, protein-RNA interactions often require RNA to fold into specific 3D shapes in order to precisely spatially present relevant chemical moieties on sugar-phosphate backbone or nucleoside bases [50]. It has been recognized in many cases that the complementarity in overall shape and structure between protein and RNA is as important as simple conservation of consensus RNA sequence. Just like in protein-protein interactions, in addition to static structures, intrinsic dynamics of RNA molecules also plays important roles in interaction between RNA and other molecules [51]. Towards this goal, we performed large scale coarse-grained molecular dynamics simulations to model the 3D structure of four well-studied RNA binding proteins (G3BP1, RBM15, LIN28B, and

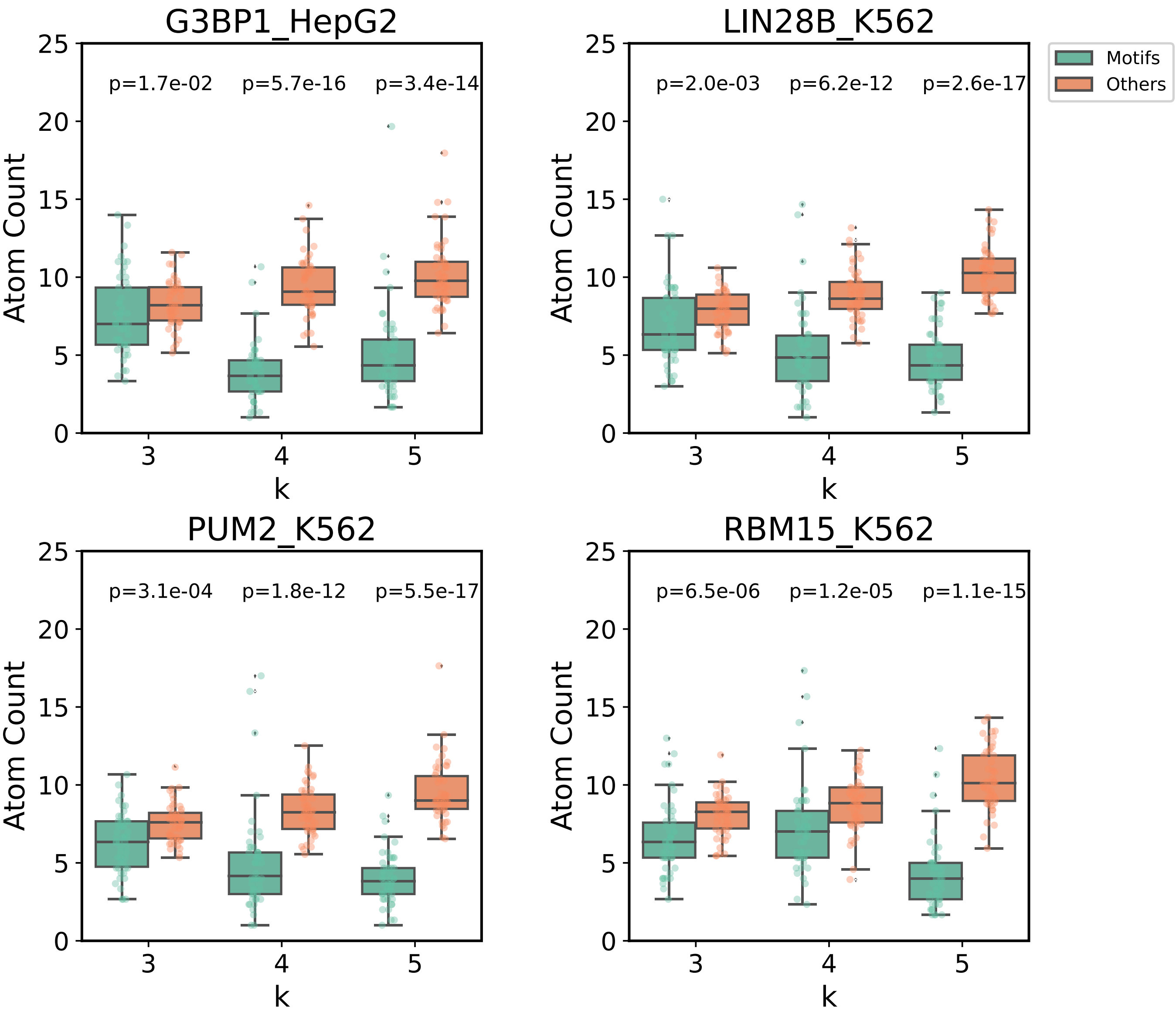

Fig 6. Boxplots of average atom counts calculated from DMD modeled 3D structures of discovered RNA network motifs and other regions of the RNA.

https://doi.org/10.1371/journal.pcbi.1010293.g006

PUM2). In addition to the RNA dynamics study, for LIN28B, we also performed protein-RNA docking and all-atom molecular dynamics simulations to validate our predicted motifs.

Following the protocol described in **Methods** (Section **4.7**), we modeled the structure of bound RNA molecules for the four aforementioned RBPs. To achieve effective binding between protein and RNA, the relevant RNA moieties need to be exposed and accessible to amino acid residues on the RBP. We constructed coarse-grain structures of RNAs and calculated spatial accessibility of the predicted RNA structural motifs by counting the number of neighboring atoms within a 15 Å radius. Our hypothesis is that these structural motifs would have higher accessibility than the rest of the RNA molecules. For the four well-studied RBPs (G3BP1, LIN28B, PUM2, and RBM15), we modelled 50 sequences for each RBP with 3 replica per sequence and performed one-sided Mann Whitney U Test (Wilcoxon Rank Sum Test) to compare the average atom counts of the identified motifs with all other k-mer regions. Fig 6 shows that for all of these RBPs, the identified RNA structure motifs have statistically significant fewer atoms in their vicinity than the rest of the RNA molecule, indicating higher spatial accessibility for these motifs. As an example, sequence and structure model of top representative motifs of each RBP are shown and the predicted EKS motifs are highlighted in **S1–S4 Figs**. Similarly significant results modelled by SimRNA software are shown in **S5 Fig**.

2.5 Case study: LIN28B-RNA docking and MD simulation

We next performed protein-RNA docking and simulation on LIN28 to further test the role of the predicted RNA structure motif in RBP-RNA recognition. We selected LIN28 for this purpose since this protein is well studied and there exist multiple high-resolution structure of it complexed with single stranded RNA [52,53].

We chose two representative sequences from the 5mer motifs and one from 4mer motifs to perform all-atom structure refinement and docking (see **Methods 4.7**). We chose the ClusPro web server to perform RNA-protein docking, since it uses an FFT based method that can do a quick, accurate, and unbiased global search to explore protein-RNA binding modes [54]. Interestingly, in all the tested cases, the binding poses with the lowest energy all captured the loop 4 in LIN28, which is a highly charged hairpin loop at the tip of the 4–5 beta sheet corresponding to residues 88–95. As shown in **Fig 7**, this loop is inserted into the motifs predicted by RNANetMotif. Subsequent MD simulations confirmed that this binding mode is energetically stable and strong phosphate contacts are formed while additional base contacts also formed, mainly through three Lys residues on this loop (Lys 88, 89, 92) (S1 Table and **Fig 7**). In the case of 4mer simulation, this loop-RNA contact served as an anchor and additional contacts were formed after certain RNA conformational changes. In a recent study, NMR and

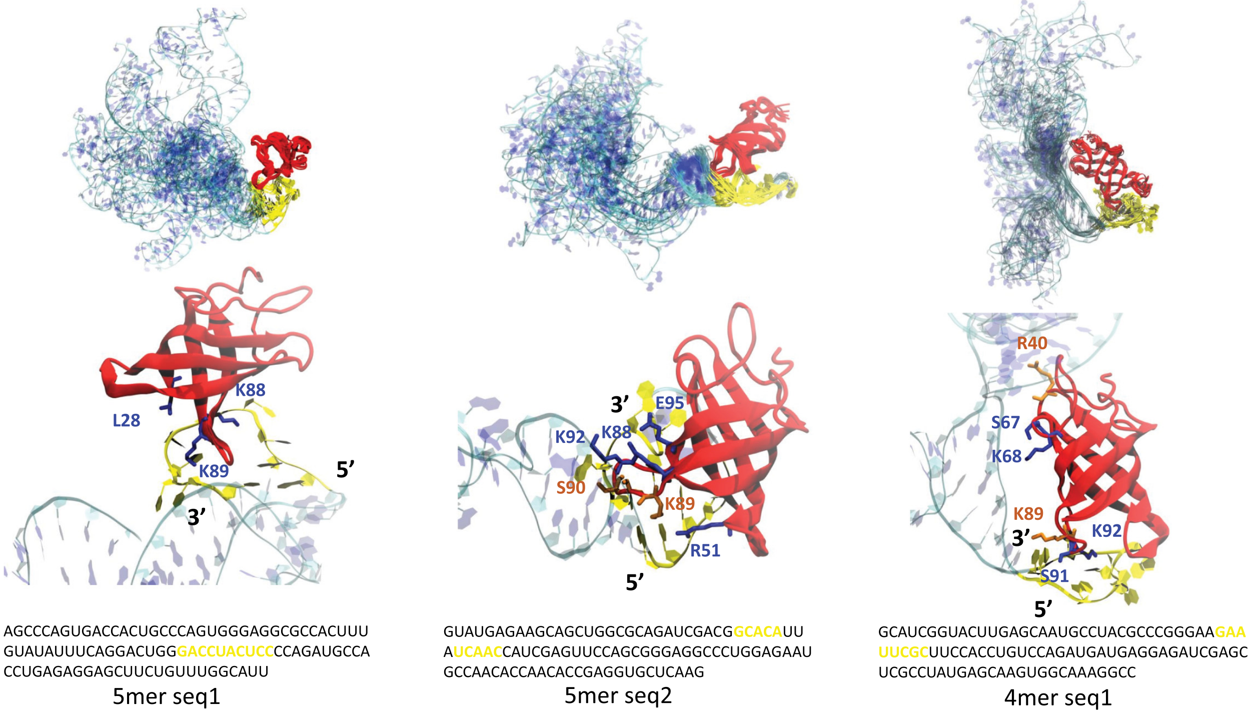
K88
L28
K89
5’
3’
3’
E95
K92
K88
S90
K89
R51
5’
R40
S67
K68
K89
3’
K92
S91
5’
AGCCCAGUGACCACUGCCCAGUGGGAGGCGCCACUUU
GUAUAUUUCAGGACUGGGACCUACUCCCCAGAUGCCA
CCUGAGAGGAGCUUCUGUUUGGCAUU
5mer seq1
GUAUGAGAAGCAGCUGGCGCAGAUCGACGGCACAUU
AUCAACCAUCGAGUUCCAGCGGGAGGCCCUGGAGAAU
GCCAACACCAACACCGAGGUGCUCAAG
5mer seq2
GCAUCGGUACUUGAGCAAUGCCUACGCCCGGGAAGAA
UUCGCUUCCACCUGUCCAGAUGAUGAGGAGAUCGAGC
UCGCCUAUGAGCAAGUGGCAAAGGCC
4mer seq1

Fig 7. Complex simulation ensemble of docking of LIN28B CSD domain and three RNA network motifs. MD ensembles and snapshot of the protein-RNA interfaces are show at the top and bottom respectively. LIN28B CSD protein is shown in red and 100 nt RNA shown in cyan and blue. The identified RNA network motifs are shown in yellow in both structure and in sequence.

https://doi.org/10.1371/journal.pcbi.1010293.g007

mutation studies indicated that K88, K89 (K98, K99 in LIN28A) play a role in electrostatic interactions with nucleic acids [55], and another recent study found a highly conserved Lys residue in YB1 protein (equivalent to K92 in Lin28B) is involved in ssDNA binding [56], as did the structural study of LIN28 on ssRNA nucleotides [57]. Altogether, these evidence suggest that our graph-based approach is very effective in finding structure elements in RBP binding sites that are important in RBP-RNA recognition. To the best of our knowledge, such a unique approach has not been described in the literature.

2.6 Comparison with other sequence-structure motif predictors on eCLIP datasets

We next compared RNANetMotif with three other computational methods, BEAM, GraphProt and ssHMM on the eCLIP datasets; two different secondary structure predictors were shown for ssHMM. We displayed the motifs identified by RNANetMotif and by other methods in **Fig 8** and summarized the common motif elements in the right column. As seen in **Fig 8**, motifs identified by RNANetMotif closely resemble the motifs predicted by other methods on eCLIP datasets, for example, PUM2 recognizes U-rich internal loop and UCHL5 recognizes AGAA in the multiloop region.

In addition, we used RNANetMotif as a supervised classifier and compared it with GraphProt on 16 RBPs. To evaluate the performance of the two classifiers for a specific RBP, we used the binding sites of the RBP as positive set and took binding sites of another RBP with similar nucleotide frequencies as the negative test set (see Text A in S1 Method and **S6 Fig**). We randomly divided the positive and negative sets into a training set and a test set using a ratio of 8:2. The test set is formed by the positive test set and the negative test set with a ratio of 1:1. As shown in **S7 Fig**, the ROC curves indicated that RNANetMotif had robust and better performance than GraphProt, as RNANetMotif had higher AUROC values than GraphProt on 14 out of 16 RBPs.

2.7 Recovery of sequence and structure motifs from synthetic datasets

Having demonstrated the effectiveness of RNANetMotif in finding important RNA structural elements in RBP bound RNA sequences, we next investigated whether this approach is also effective in finding meaningful RNA structural motifs in a more general setting. As shown in **Table 1,** we selected five RNA sequence-structure motifs of representative types, i.e., hairpin loop, bulge, internal loop, hairpin loop and multi loop. We then randomized the flanking nucleotide sequences (represented as Ns) while maintaining the structure motif. We synthesized 100 instances of these sequence motifs and inserted each instance into a randomly generated RNA sequence of 100 nt long. For each type of motif, we also created a set of random background sequence set of the same size as the negative control set. We next ran RNANetMotif and three other methods, GraphProt, RNAcontext and Zagros, on these sequence sets to try to recover these spiked-in motifs. As shown in **Table 2**, RNANetMotif consistently recovered the sequence and structure of every one of the five motifs, which compares favorably with other methods. The detailed recovering results of these four methods are summarized in **Table 2.**

2.8 Run-time and memory comparisons

We compared the run time of RNANetMotif with other methods on five eCLIP datasets (see **S8 Fig**) and model training time for GraphProt. RNANetMotif consumed the least time among all the five tools, while GraphProt ran a little slower than RNANetMotif on three datasets and faster than RNANetMotif on two other datasets. All the tools had similar level of

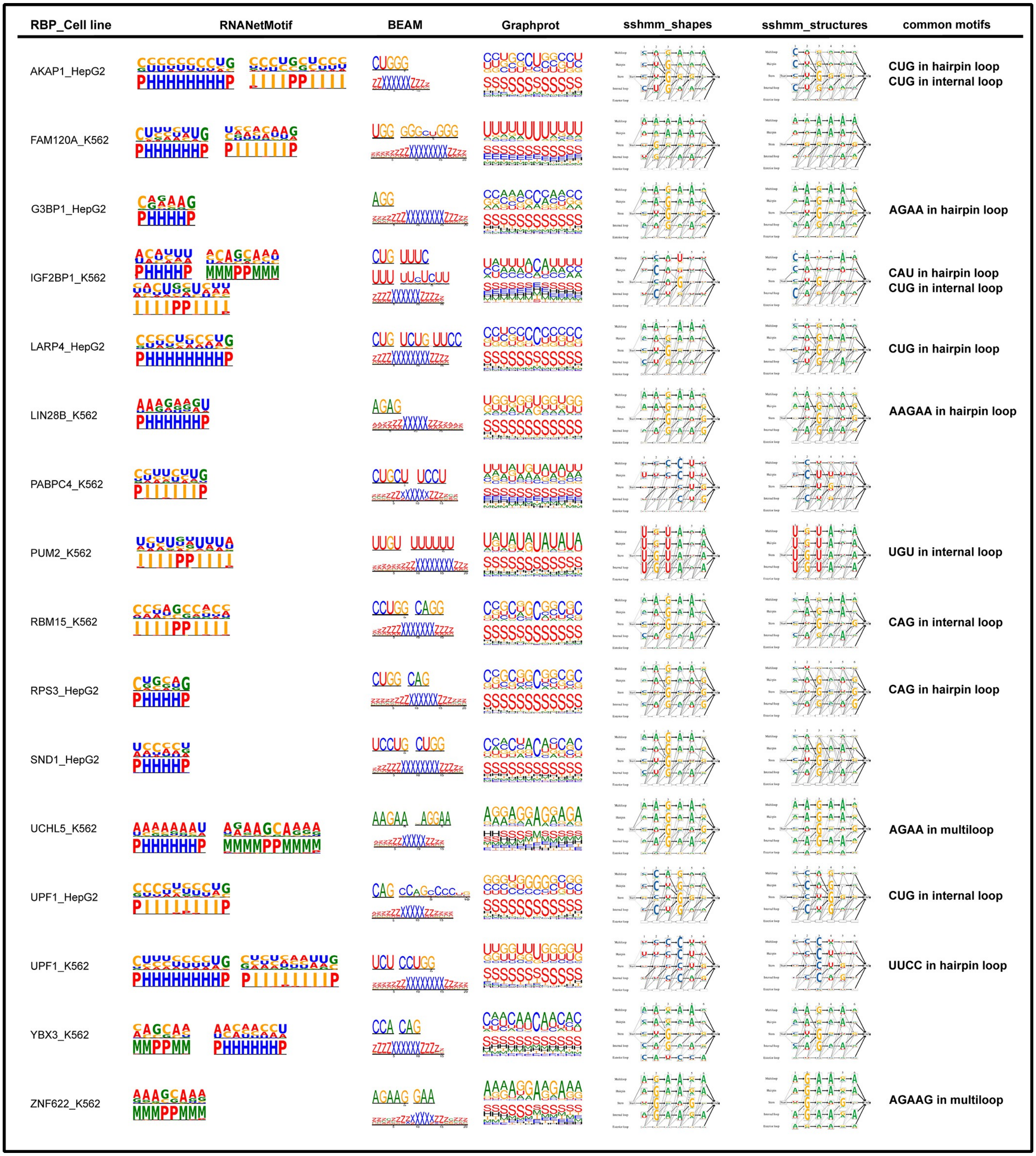

Fig 8. Comparison with other sequence-structure motif predictors on 16 eCLIP datasets.

https://doi.org/10.1371/journal.pcbi.1010293.g008

memory usage, with maximum memory usage of no more than 10 GB on all the tested data. We like to note that a more comprehensive and robust investigation and benchmark is warranted to test the utility of RNANetMotif with regards to this task.

Table 1. Implanted RNA motifs in recovery analysis as described in Section 2.7.

Implanted Motifs		
CCACCA in hairpin loop	NNNCCACCANNN	Sequence
	(((.. ....)))	Structure
AUG in bulge	NNNAUGNNNNNNNNNNNNNNNN	Sequence
	(((.. . (((.. .... ..)))))))))	Structure
AA_AA in internal loop	NNNAANNNNNNAANNN	Sequence
	(((.. ((())) ..)))	Structure
GAGAGAGA in hairpin loop	NNNNNGAGAGAGANNNNN	Sequence
	(((((.. .... ..)))))	Structure
CCCC_AAAA_CCCC in multiloop	NNNCCCCNNNNNNAAAANNNNNNCCCCNNN	Sequence
	(((.. .. ((((())))) .. .. (((((((()))))))) .. ..)))	Structure

https://doi.org/10.1371/journal.pcbi.1010293.t001

3. Discussion

In this paper we describe a novel network-based approach in finding meaningful sequence and structure motifs from RBP bound RNA sequences. We recognize that there have been many other methods developed over the years that have addressed this problem from many different angles (reviewed in **Introduction**). The novelty of RNANetMotif is that it takes a graph-based approach to extract enriched subgraphs from RNA secondary structures. Ideally, the most accurate and unbiased way to determine the binding mechanism between an RBP and its target RNAs is to compare high-resolution 3D structures of a representative set of RBP-RNA complexes and derive a set of common 3D structure elements shared by these structures. However, this is technically challenging and logistically unrealistic. It has been well documented that linear representations such as Positional Weight Matrices (PWM) have their limitations in capturing the binding preferences [58,59]. Several methods such as BEAM or GraphProt have improved upon PWMs by adding structural descriptors to each position, i.e., helix, stem, loop, which has shown improvement [16,60]. Despite these recent developments, we feel there is still room for improvement. Motivated by the observation that RNA secondary structures, represented as a network, are essentially low-resolution abstracted structure of RNA molecules, we hypothesized that there exist enriched subgraphs in these networks that are determinants of the recognition process between RBP and their RNA targets. Conceptually, this is analogous to the study and design of protein structures from the aspect of hydrogen bonds and other interactions among amino acid residues, i.e., contact maps.

We introduced a new concept to represent RNA elements, i.e., EKS (Extended K-mer Subgraph) and an algorithm GraphK. As demonstrated in this study, we think EKS is a promising approach in mining local RNA secondary structures for enriched motifs. We further showed

Table 2. Recovery rates of implanted RNA motifs by different methods (Section 2.7).

Software (type)	3nt bulge loop	4nt internal loop	6nt hairpin loop	8nt hairpin loop	12nt multi loop	Overall recovery rate
RNANetMotif (Sequence)	1	1	1	1	1	1
RNANetMotif (Structure)	1	1	1	1	1	1
GraphProt (Sequence)	1/3	1	2/3	5/8	7/12	2/3
GraphProt (Structure)	1/3	1	0	0	2/3	2/5
RNAcontext (Sequence)	1/3	3/4	2/3	1	5/6	5/7
RNAcontext (Structure)	0	0	0	1	1	2/5
Zagros (Sequence)	2/3	3/4	1	7/8	2/3	4/5
Zagros (Structure)	0	1	1	1	2/3	3/4

https://doi.org/10.1371/journal.pcbi.1010293.t002

that it can be extended to the study of more general problems in addition to RBP-RNA interactions, although more rigorous benchmarking is required. Compared to other methods, RNANetMotif has the following unique characteristics. (i) It recognizes RNA target sequences as networks and applies graph theory approaches to extract meaningful and enriched subgraphs. (ii) In contrast to other methods, RNANetMotif follows an unsupervised scheme thus avoids the noise and uncertainty introduced by the false negative data. (iii) We also conducted structure modeling and molecular dynamics studies and validated some of the predictions. RNANetMotif is a new way of investigating RNA sequences and can be further extended into the analysis of other categories of RNA sequences. The source code of RNANetMotif can be accessed at Github (https://github.com/hongli-ma/RNANetMotif). We also constructed a website that allow users to test their own data: http://rnanetmotif.ccbr.utoronto.ca.

There are several limitations and avenues for future improvements on our current approach. When building a network of RNA secondary structure interactions, we only included canonical Watson–Crick and GU wobble base pairs and excluded other non-canonical base pairing such as Hoogsteen base pairing or Leontis-Westhof interactions [61,62]. This is mostly due to the lack of high-resolution 3D RNA structures, which makes accurate prediction of these non-canonical interactions less feasible. Newer RNA structure prediction methods such as MXfold2 use deep learning approach and have achieved promising results [63]. Deep learning methods such as DeepRiPE and PrismNet have also shown promises in modeling and predicting RBP-RNA interactions [33,64]. More intriguingly, several recent methods have adopted graphic neural networks (GNN) in modeling RNA secondary structures [34,65]. Conceptually similar to RNANetMotif, these methods represent the predicted RNA secondary structure as a network and assign features to each node and each edge according to the types of nucleotides and types of the chemical bonds. Messages encoding the embedded node and edge features are passed between neighboring nodes, and integrated gradients (IG) is used to extract discriminating RNA sequences [66]. The difference between the GNN approaches and RNANetMotif is that RNANetMotif is an unsupervised approach and searches for enriched motifs directly. Nevertheless, it would be very intriguing to combine these two types of approaches in the future. We also focused on in vivo RBP binding data in this study; it would be very interesting to extend this graph-based approach to in vitro data generated by RNAcompete or HTR-SELEX.

Lastly, the majority of the RBPs currently being investigated by iCLIP or eCLIP preferentially bind to single stranded RNAs (ssRNAs). We envision that RNANetMotif or other similar approaches such as RNAcompete or GraphProt are more amenable to the study of these ssRNA binding proteins than to dsRNA binding proteins, since single stranded RNAs have a more diverse range of secondary structure elements and these elements are presumably important in the binding process. On the other hand, although the binding sites of the dsRNA binding proteins are double stranded stem structures, it is also possible that other regions outside of the binding sies have enriched structure elements that put constraint on the RNA and possibly help present the stem structure to the incoming RBPs. A more detailed examination of these potential structure elements is warranted when more eCLIP data become available for these dsRNA binding proteins.

4. Materials and methods

4.1 Collection of RBP eCLIP data and preprocessing

We downloaded the "enhanced UV crosslinking followed by immunoprecipitation" (eCLIP) data (1) from ENCODE project website (release v101) (https://www.encodeproject.org/) in June 2020, which consists of chromosomal peak regions that are bound by specific RBPs. The peaks were annotated as IDR BED files which had been processed following the eCLIP-seq Processing Pipeline. Reproducible and significant peaks that passed the Irreproducible Discovery Rate (IDR) were identified by a modified IDR method.

A total of 223 IDR files corresponding to 150 unique RNA binding proteins (RBPs) were collected, which comprised of 120 files for K562 cell line, 103 files for HepG2 cell line. We next retained only 56 IDR files that have more than 5,000 mapped binding sites. We further excluded those proteins that were annotated to be primarily involved in mRNA splicing and excluded those RBPs that had fewer than half of the binding sites mapped to the annotated exon regions. The final derived set contained 22 IDR files. The rational of above procedures is to limit our study only to those RBPs that primarily bind to the mature mRNA transcript region. To unambiguously extract exon regions, we used the most prominent transcript for each gene as was defined based on the basic gene annotation from GENCODE (Release 34, GRCh38.p13) through hierarchical filtering: first filtered by APPRIS annotation (highest priority) [67], then by transcript support level, and finally by transcript length (longer isoform preferred). We only kept one prominent transcript per gene.

We next processed the 22 IDR files by only keeping the peaks that are entirely localized in one of the exons of the prominent transcript., i.e., removing those binding sites in the introns. The detailed information of the 22 RBPs with specific cell line is summarized in **Table 3.** We chose 100 nt as the uniform binding site length (50 nt extensions up- and down- stream of the center position) and used ’getfasta’ function from ’bedtools’ to map BED files to FASTA files. We then removed redundant and missing data in the binding sites by using module ’cd-hit-est’ from CD-HIT at 80% similarity cut-off [68]. A list of global tunable parameters used in RNANetMotif is listed in Text B in S1 Method.

Table 3. RBPs and the domain information.

RBP and Cell Line	RNA binding domain	#peaks
AKAP1_HepG2	KH,Tudor	5338
BCLAF1_HepG2		22884
DDX24_K562	Helicase ATP-binding, Helicase C-terminal	5841
DDX3X_HepG2	Helicase ATP-binding, Helicase C-terminal	5976
DDX3X_K562	Helicase ATP-binding, Helicase C-terminal	3961
FAM120A_K562		4218
G3BP1_HepG2	NTF2, RRM	5204
GRWD1_HepG2		16040
IGF2BP1_K562	RRM1, RRM2, KH1, KH2, KH3, KH4	4690
LARP4_HepG2	HTH La-type RNA-binding, RRM	4350
LIN28B_K562	CSD	4616
PABPC4_K562	RRM1, RRM2, RRM3, RRM4, PABC	4865
PPIG_HepG2	PPIase cyclophilin-type	13538
PUM2_K562	PUM-HD	4742
RBM15_K562	RRM1, RRM2, RRM3, SPOC	6800
RPS3_HepG2	KH type-2	4697
SND1_HepG2	TNase-like 1, TNase-like 2, TNase-like 3, TNase-like 4, Tudor	5697
UCHL5_K562		12866
UPF1_HepG2		8547
UPF1_K562		11708
YBX3_K562	CSD	12706
ZNF622_K562		10551

https://doi.org/10.1371/journal.pcbi.1010293.t003

4.2 Predicting RNA secondary structures and graph construction

We used the RNAplfold program in the ViennaRNA package (version 2.4.13) for secondary structure prediction. The stem candidates were derived from the base pairing probability matrix calculated by McCaskill's algorithm (43). After experimenting with the choices of window sizes (L and W-L), we set the parameters of RNAplfold as W = 100, L = 100 while allowing Watson-Crick (A:U and G:C) and wobble G:U base pairing [69]. We note that, instead of generating a final structure, RNAplfold calculates base-pairing probabilities between nucleotide pairs in the entire sequence. For each RNA sequence, we only keep those reliable base pairs with probability >0.5, which do not form a pseudoknot and are in a centroid [70].

Having predicted base-pairing probabilities in the entire RNA, we next predicted and refined local secondary structure probabilities associated with each nucleotide by using the software RCK [20]. The following five type representations are considered: H for hairpin loop, I for internal loop, M for multi-loop, E for external loop, and P for paired. The following parameters were used: W = 100, L = 100, u = 1. The final combined features for each nucleotide consist of its nucleotide type and the probabilities of having each of the five secondary structures, i.e., H, I, M, E and P.

We represented each 100 nucleotides (nt) long RBP bound RNA sequence as a weighted graph G = (V, E, w), where V consists of nucleotides encoded as vertices with discrete labels (A, C, G, U) and continuous labels (see above). The edge set E contains RNA backbones and the predicted base-pairings. The weight of backbone edges is defined as 1, while base-pair edges (Watson-Crick or G: U) as the probability predicted from RNAplfold.

4.3 Definition of EKS–Extended K-mer Subgraph

In this work, we introduce a new concept dealing with graph representation of local RNA secondary structures, referred to as Extended K-mer Subgraph or EKS. Given two base-pairing nucleotides, *i* and *j*, we build a subgraph by extending along the backbone edges from *i* and from *j*, separately and respectively (**Fig 2**). During the extension process, we traverse along the backbone from *i* and from *j* respectively, until a k-mer is reached in both directions. We define the resultant subgraph containing the *2k* number of nucleotides as an Extended k-mer Subgraph (EKS). The intuition of EKS is that these subgraphs represent smaller tractable network elements as component of a larger complex network.

As illustrated in **Fig 2**, the linear sequence of the local folded RNA follows the following direction: $5' \to s \to s' \to t' \to t \to 3'$. As an example, given k = 3, we select base pair (i, j) as a root base pair, and denote U = {i, j} as the initial vertex set. We next perform k-mer extension process along the backbone edges, and extend U by adding nodes from the k-mer containing node i and the k-mer containing node j. After this process, the number of nodes in U reaches 2k and we denote it as G[U], the subgraph of G induced by U, as an EKS. The notation G[U] represents the subgraphs of G with vertex set U and all the weighted edges that have both end vertices in U. We define the diameter of the graph as the "longest shortest path" between any two vertices in the graph. We also define the "compactness score "of G[U] as the reciprocal of the diameter of G[U], which is calculated by Floyd-Warshall algorithm with the time complexity of $O(|U|^3)$ [71, 72]. The benefit of using compactness score is that it falls in the range of (0, 1) thus is easier to manipulate in scoring or probability functions. Lower compactness score is indicative of lower "connectiveness" of the network, i.e., existence of node pairs separated by longer path.

Depending on the directions of traversing from the root base pair (i, j), we can divide the set of EKSes into three categories: opposite-direction extensions, mixed-mode extensions, and same-direction extensions. As shown at the top of **Fig 2**, the opposite-direction extensions can generate EKSes that are either left-opened or right-opened subgraphs. Mixed-mode extensions have the most diverse shapes and topologies. As for the same-direction extensions, the network is extended in the same direction, both toward 3' end or 5' end. The nucleotides in these EKSes have lower chance of forming base-pairing unless pseudoknot structure is allowed.

4.4 GraphK–a novel subgraph extraction approach

Having defined EKS, we next introduce an algorithm, GraphK, which applies extension, extraction, and filtering steps to partition a complex graph into overlapping representative EKSes. The detailed steps of the GraphK algorithm are as follows **(Fig 1A)**. In **Step 1**, we conduct k-mer extension as described above in each RBP-bound RNA sequence to obtain all EKSes associated with the RNA sequence. In addition, we add an annotation to each EKS, including its type, compactness score, and the starting positions of the two k-mers. In **Step 2,** we extract EKSes generated from opposite-direction extensions (i.e., right-opened, or left-opened), and exclude those generated by same-direction or mixed-mode extensions. The rationale is to extract only those likely to form single stranded RNAs (see above). In **Step 3,** we filter EKSes by their compactness scores and only retain those with lower compactness scores, i.e., those with less base pairings. We next filter the EKSes by the position of the nucleotides and keep the EKSes whose sequence is entirely located within the central 40 nucleotides. By this approach, we partitioned every 100 nt RNA into local network elements, further filtered and pooled them into the final EKS pool.

In the following, we elaborate the rationales behind the GraphK algorithm. First, it is difficult to represent the 2D or 3D structure of the entire RBP binding region. Therefore, we treated the predicted secondary structure of the entire RBP binding region as a network and extracted base-pair derived subgraphs from the network to represent components of the entire structure network. These local subgraphs are more tractable than a global graph in representing the entire RNA region. Our hypothesis is that the RBP-bound RNAs could share some of these common subgraphs. Secondly, we briefly explain the choice of size, shape, and compactness score of the EKS. Since most single RNA binding domains (RBDs) appear to bind RNA motifs 3 ~ 8 nt long [12], we chose k ranging from 3 to 5, corresponding to 2k ranging from 6 to 10. As a majority of the RBPs we studied prefer single stranded RNA, we only considered the subgraphs created by opposite-direction expansion which generates single stranded RNAs (see **Fig 2**). In addition, we used a graph theory method to calculate compactness scores for EKSes and retained those with lower compactness score, i.e. those EKSes with sparse base-pairs and are more prone to binding by RBPs. Thirdly, we explain the choice of 40 nucleotides as the length of EKS. This is because the mean length of the peaks from eCLIP datasets is around 40 ~ 60 nt and the structural distribution in the central regions of peaks shows better structural conservation.

4.5 Calculate HVDM distance matrix and construct EKS similarity network

We next calculate similarities between any pairs of EKSes so that we can select representative and discriminative EKSes associated with each RBP binding region. Each instance of EKS consists of $2k$ nucleotides and has $12k$ mixed-type attributes, comprised of $2k$ categorical attributes (i.e., nucleotide types as A, C, G, U) and $10k$ continuous attributes (i.e., secondary structure probabilities). We chose to use the heterogeneous value difference metric (HVDM) as the distance function, which uses value difference metric (VDM) to handle categorical attributes, and absolute differences to handle continuous attributes [45,73]. Given two EKSes with the feature vector $\overrightarrow{x} = (x_i, i = 1, 2, \ldots 12k)$ and $\overrightarrow{y} = (y_i, i = 1, 2, \ldots 12k)$, in the following we describe the details on the calculation of HVDM distance.

First, we classify an EKS according to which region on the RNA sequence the first and last nucleotide of the EKS falls into and assign the EKS to one of the following classes: *classes* = {C_i, i = 1,2,3,. . .,10}. As shown in **S9 Fig**, we first partition the central 40 nucleotides into 4 intervals of equal length of 10 nt as {I_k, k = 1,2,3,4} and assign the EKS to one of the following classes according to the intervals into which the first and the last of the nucleotide falls: $C_i \in \{I_{i,j}$, i, j = 1,2,3,4. $i \leq j\}$. The VDM distance between $\overrightarrow{x}$ and $\overrightarrow{y}$ is defined in Eq (1), which considers the correlations between each possible attribute value and each EKS class.

$$vdm_{\alpha}\left(\overrightarrow{x},\overrightarrow{y}\right) = \sqrt{\sum\nolimits_{h=1}^{\#classes}\left|P(C_h|x_{\alpha}) - P(C_h|y_{\alpha})\right|^2} = \sqrt{\sum\nolimits_{h=1}^{\#classes}\left|\frac{N_{\alpha,x_{\alpha},C_h}}{N_{\alpha,x_{\alpha}}} - \frac{N_{\alpha,y_{\alpha},C_h}}{N_{\alpha,y_{\alpha}}}\right|^2} \tag{1}$$

where

$$N_{\alpha,x_{\alpha}} = \sum\nolimits_{h=1}^{\#classes} N_{\alpha,x_{\alpha},C_h} \tag{2}$$

$$P(C_h|x_{\alpha}) = \frac{N_{\alpha,x_{\alpha},C_h}}{N_{\alpha,x_{\alpha}}} \tag{3}$$

$N_{\alpha,x_{\alpha},C_h}$ is the number of instances in all EKSes that have value x_{α} for attribute α and class C_h; $N_{\alpha,x}$ is the number of instances in all EKSes that have value x_{α} for attribute α, i.e., $N_{\alpha,x_{\alpha}}$ is the sum of $N_{\alpha,x_{\alpha},C_h}$ over all classes. $P(C_h|x_{\alpha})$ is the conditional probability that the class is C_h given that attribute α has the value x_{α}. The sum of $P(C_h|x_{\alpha})$ over all classes is 1 for a fixed value of α and x_{α}.

We next define the difference between two values x_{β} and y_{β} of a continuous attribute β as the absolute difference:

$$diff_{\beta}(\overrightarrow{x},\overrightarrow{y}) = |x_{\beta} - y_{\beta}| \tag{4}$$

Taking these together, we then define the HVDM distance between $\overrightarrow{x}$ and $\overrightarrow{y}$ as:

$$\mathrm{HVDM}(\overrightarrow{x},\overrightarrow{y}) = \sqrt{\sum\nolimits_{a=1}^{12k} {d_a}^2(x_a, y_a)} \tag{5}$$

where

$$d_a(x_a, y_a) = \begin{cases} vdm_a(\overrightarrow{x},\overrightarrow{y}), \textit{if a is categorical} \\ diff_a(\overrightarrow{x},\overrightarrow{y}), \textit{if a is continuous} \\ 1, \textit{other conditions} \end{cases} \tag{6}$$

Next, we initialized a weighted complete graph S (V_S, E_S, W_S) with EKSes as vertices and the above calculated HVDM distance between vertices as the weight of the edges. To denoise this graph, we then refined this graph by the following pruning steps. Firstly, we selected those vertices that are close to each other and have similar partners on the network. For v_1, $v_2 \in V_S$, if v_1 and v_2 are among top n nearest neighbors of each other, i.e., v_1 is one of the top n nearest neighbors of v_2 and v_2 is also one of the top n nearest neighbors of v_1, v_1, v_2 and the edge between them (i.e., e_{v_1,v_2}) are retained. We then removed the remaining vertices and edges. By doing so, we obtained a vertex-pruned network and each pair of nodes in it is in a very close relationship with a short HVDM distance. To reveal the relationship between vertices in the vertex-pruned network more discriminatively, we replaced the distance-based similarity with neighborhood-based similarity and re-defined the weight of the edges e_{v_1,v_2} between nodes v_1 and v_2 in this network as the Jaccard index in Eq (7). N_{v_1} and N_{v_2} are the neighbor sets of v_1

and v_2 respectively.

$$W_{e_{v_1,v_2}} = Jaccard\left(N_{v_1}, N_{v_2}\right) = \frac{N_{v_1} \cap N_{v_2}}{N_{v_1} \cup N_{v_2}} \tag{7}$$

In the second pruning step, we removed any edges e with weight lower than a cutoff, $w_e \leq w_{cut\text{-}off}$; the purpose of this was to retain only those sufficiently similar EKS candidates. After experimenting with different parameters, we set n = 2% × *size of* V_S, w_{cutoff} = 0.9 *quantile of all weights*. We denote the pruned graph as the similarity network among filtered EKSes. In **S2 Table**, we summarize the number of the nodes and edges of the similarity network for the 22 RBPs.

4.6 Overall binding preference and intrinsic network binding motifs

Having constructed the similarity network among EKSes for each RBP, we next took a multi-steps approach to identify EKSes that are important in RBP binding. We first searched the similarity network for densely connected local modules, which is analogous to searching for protein complexes in protein-protein interaction networks. We retained those significant local modules by comparing them with a negative control set. We then searched these significant modules for network cliques; those EKSes present in overlapping cliques are deemed as important for RBP binding. Analogous to the study of protein-protein interactions, this is akin to finding proteins that are "hub" nodes on a protein-protein interaction network.

We used the ClusterONE software [46] to derive densely connected modules; the parameters were set at (-s 100, -d 0.8,—max-overlap 0.2), i.e. requiring modules having density higher than 0.8 and a size of at least 100 nodes. The purpose of this step is to ensure the similarity network among EKSes we constructed is sufficiently dense to conduct maximal clique detections. To evaluate the significance of network modules of a specific RBP, we generated negative sites from non-target RBP binding datasets using a greedy strategy to ensure that both the positive set and the negative set have the similar single and di- nucleotide frequencies (see Text A in S1 Method for details). We then compared Positional Weight Matrices (PWMs) of modules in the same structural context of these two sets by using Pearson χ^2 test [74]. We note that the χ^2 p-value which is calculated for each column in PWM is based on the null hypothesis that the aligned columns are independent and identically distributed observations from the same multinomial distribution. We used the geometric mean of the column p-values as the p-value of module for the significance evaluation.

After confirming the existence of dense modules in the similarity network, we next applied Bron–Kerbosch algorithm [75–77] for maximal clique enumeration (MCE) over the similarity network to obtain the maximal cliques; we retained the upper half of the maximal cliques according to the rank of their sizes. A maximal clique is defined as a clique that cannot be extended by inclusion of additional adjacent vertices, i.e., it is not a subset of a larger clique. The objective of this step is to identify representative EKS or subgraph in a densely connected region in a similarity network so that it can capture the modality of the network. In practice, maximal cliques in the EKS similarity network are highly overlapping, in other words, an EKS can be part of multiple maximal cliques. These EKSes are likely representative of the overall topology of the dense subnetworks and the binding preference of the RBP, subsequently we extracted the top 50 of these recurring EKSes and considered them as RBP structural element that are important for RBP binding.

4.7 RNA 3D structure modeling and evaluation of network motifs

The RNA 3D structures were predicted using a discrete molecular dynamics (DMD) method with predicted base-pair information provided as constraints [44]. The DMD engine and scripts were adopted from the literature [78] and kindly provided by Dr. Feng Ding (http://

dlab.clemson.edu/). Starting from an RNA sequence and base-pairing information, the default DMD settings were used to perform coarse-grained folding simulation for 100,000 DMD time units; for each sequence, 3 replica of MD simulation was performed. For selected LIN28B systems, the last snapshots of DMD simulations were extracted to recover the 3-beads coarse-grained representation back to full atoms to perform more accurate full-atom docking and simulations. Since this conversion may result in severe clashes or abnormal bond length, some minimization and equilibration, as well as a short 10-ns full atom simulations were performed to further refine the structures before docking. For spatial accessibility, a 15Å cut-off was used to count the neighboring atoms of the given regions to calculate the exposure level, while nucleotides within 2 positions on both 5' and 3' directions were excluded since they would always be counted due to bonded connections except for at both ends of the sequences. To evaluate the significance of the discovered network motifs, Mann-Whitney-Wilcoxon (MWW) rank-sum U Test was performed on such neighboring counts of identified motifs against the average of all the k-mer regions. To avoid possible bias introduced by the selection of MD protocol, we also employed SimRNA [79] to model the RNA structure in coarse-grained representation with 3 replica per sequence. Different from DMD, SimRNA used 5 beads to represent one nucleotide.

Charmm-gui webserver was used to set up the MD systems [80], the complex was solvated in a water box with 15 Å buffer of water extending from the RNA or protein-RNA complex. K^+ and Cl^{2-} ions were added to ensure a 0.15M ionic concentration and zero net charge. Due to the diversity and intrinsic flexibility of RNA molecules, the built systems contain a wide range of ~300,000 to 700,000 atoms. After 10,000 steps of minimization and equilibration, where harmonic restraints were applied to heavy atoms, production MD simulations were performed in the NPT ensemble. The Nosé-Hoover method was used with temperature T = 30˚C [81]. The Parrinello–Rahman method was used for pressure coupling [82]. A 10-Å switching distance and a 12-Å cutoff distance were used for non-bonded interactions. The particle mesh Ewald method was used for electrostatics calculations [83]. The LINCS algorithm was used to constrain the hydrogen containing bond lengths for a 2-fs time step [84]. The energy minimization and MD simulations were performed by using the GROMACS program [85] (version 2019.6-CPU/GPU) using the CHARMM36m force field [86–88] and TIP3P water model [89].

The RNA structures after 10-ns production run were used as receptors and the LIN28B CSD structure (PDB ID 4A4I) [57] was used as ligand to perform protein-RNA docking using the ClusPro webserver [54,90]; the predicted RNA motifs was provided as binding site constraints. The most populated docking poses were subjected to another 100 ns MD simulation to further test the binding stability; the last 50 ns equilibrated trajectories were used for analysis and visualization (**S10 Fig**). We used the following geometric criteria to identify a hydrogen bond (HB) between two polar non-hydrogen atoms (i.e., acceptor and donor): the donor-acceptor distance is <3.5 Å, and the deviation of the donor-hydrogen-acceptor angle from 180˚ is <60˚. We used the VMD program to identify and calculate the occupancy of HBs, the root-mean-square deviation (RMSD) of the protein and motif and for visualization {Humphrey, 1996 #10}.

4.8 Robustness of our approach with different parameters

It is known that base-pairing probabilities predicted by RNAplfold or any other software can be influenced by global parameters such as the size of the window (W) and the maximum base pair span (L). We experimented with different combination of W and L: (100, 100), (100, 50), (80, 40) and (70, 70), and tested on four representative RBPs, G3BP1, IGF2BP1, LIN28B and PUM2 (also see Text C in S1 Method).

S11 Fig compares the RNAplfold prediction results with these different parameters, which shows that the base-pairing probabilities largely followed the same overall shape. This suggest that, within the confine of our task at hand, the prediction results of RNAplfold are largely robust with regard to slight changes in global parameters. We also re-ran RNANetMotif pipeline with the base-pairing predictions generated from the aforementioned four sets of (W, L) parameters, the generated RNA network motifs were further used in a supervised classifier to separate true or false RBP binding sites. As shown in S12 Fig, the overall trend and the shape of the ROC curves are very similar to each other.

We also compared the prediction results between RNAplfold with several other prediction tools. We compared the results of RNAplfold, RNAstructure, RNAfold and calculated the Jaccard similarity of the predicted base-pairs (S13 Fig), which showed general consistency among these three methods. We also compared in silico predicted RNA base-pairing probabilities with the experimentally determined RNA nucleotide accessibilities, i.e., the icSHAPE scores [91]; the results are also consistently congruent, indicating the validity and robustness of the computational approaches.

Supporting information

S1 Table. Hydrogen bond table based on RNA-Protein complex MD simulation.
(XLSX)

S2 Table. Summary of the number of the nodes and edges of the similarity network for the 22 RBPs.
(XLSX)

S1 Method. A. Selection of control sites for the evaluation of network modules. **B.** Summary of parameters in RNANetMotif algorithm. **C.** Parameters in RNA secondary structure prediction.
(DOCX)

S1 Fig. Top representative network motifs of G3BP1 for different k in 2D and 3D structures with EKS highlighted in yellow.
(PDF)

S2 Fig. Top representative network motifs of LIN28B for different k in 2D and 3D structures with EKS highlighted in yellow.
(PDF)

S3 Fig. Top representative network motifs of PUM2 for different k in 2D and 3D structures with EKS highlighted in yellow.
(PDF)

S4 Fig. Top representative network motifs of RBM15 for different k in 2D and 3D structures with EKS highlighted in yellow.
(PDF)

S5 Fig. Boxplot of average atom counts calculated from SimRNA modeled 3D structures of the network motifs and other regions for 4 RBPs.
(PDF)

S6 Fig. Nucleotide distribution in binding sites of 22 RBPs.
(PDF)

S7 Fig. Comparison between RNANetMotif and GraphProt on 16 RBPs (ROC curves).
(PDF)

S8 Fig. Comparison on running time of different tools on five eCLIP datasets.
(PDF)

S9 Fig. A schematic display of folded and unfolded RNA sequence to describe class choices in defining categorical attributes in HVDM distance. Three EKSes are marked as red, green, and brown dashed circles in the RNA secondary structure. After unfolding, these EKSes are mapped back to primary RNA as gapped sequences with two k-mers falling into four intervals $\{I_k, k = 1,2,3,4\}$.
(PDF)

S10 Fig. RMSD plot of molecular dynamics trajectories.
(PDF)

S11 Fig. Line plots comparing the prediction results of RNAplfold with different parameters. The value of Y axis stands for the trimmed mean of unpaired probability of each position
(PDF)

S12 Fig. Comparison between RNANetMotif prediction on different (W, L) values of RNAplfold for 4 RBPs (ROC curves).
(PDF)

S13 Fig. Violin plots comparing the prediction results of RNAplfold, RNAstructure, and RNAfold on five RBPs.
(PDF)

Acknowledgments

Computations in this work were performed on the Niagara Supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation; the Government of Ontario; Ontario Research Fund—Research Excellence; and the University of Toronto. HW thanks Dr. Philip Kim from Donnelly Centre, University of Toronto for his support in the early stage of this work. We thank Dr. Feng Ding and Zhenzhen Zhang from Clemson University for their helpful discussion about DMD method.

Author Contributions

Conceptualization: Hongli Ma, Zhaolei Zhang.

Data curation: Hongli Ma.

Formal analysis: Hongli Ma.

Funding acquisition: Guojun Li, Zhaolei Zhang.

Investigation: Hongli Ma, Han Wen, Zhaolei Zhang.

Methodology: Hongli Ma, Han Wen, Zhaolei Zhang.

Software: Hongli Ma, Zhiyuan Xue.

Supervision: Guojun Li, Zhaolei Zhang.

Visualization: Zhiyuan Xue.

Writing – original draft: Hongli Ma, Zhaolei Zhang.

Writing – review & editing: Hongli Ma, Han Wen, Zhaolei Zhang.

References

1. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020; 583(7818):711–9. https://doi.org/10.1038/s41586-020-2077-3 PMID: 32728246
2. Neelamraju Y, Hashemikhabir S, Janga SC. The human RBPome: from genes and proteins to human disease. J Proteomics. 2015; 127(Pt A):61–70. https://doi.org/10.1016/j.jprot.2015.04.031 PMID: 25982388
3. Matia-Gonzalez AM, Laing EE, Gerber AP. Conserved mRNA-binding proteomes in eukaryotic organisms. Nat Struct Mol Biol. 2015; 22(12):1027–33. https://doi.org/10.1038/nsmb.3128 PMID: 26595419
4. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nature Reviews Genetics. 2014; 15(12):829–45. https://doi.org/10.1038/nrg3813 PMID: 25365966
5. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell. 2012:1–14. https://doi.org/10.1016/j.cell.2012.04.031 PMID: 22658674
6. Qin H, Ni H, Liu Y, Yuan Y, Xi T, Li X, et al. RNA-binding proteins in tumor progression. J Hematol Oncol. 2020; 13(1):90. https://doi.org/10.1186/s13045-020-00927-w PMID: 32653017
7. Gebauer F, Schwarzl T, Valcarcel J, Hentze MW. RNA-binding proteins in human genetic disease. Nat Rev Genet. 2021; 22(3):185–98. https://doi.org/10.1038/s41576-020-00302-y PMID: 33235359
8. Kechavarzi B, Janga SC. Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biology. 2014; 15(1):R14. https://doi.org/10.1186/gb-2014-15-1-r14 PMID: 24410894
9. Ray D, Kazan H, Chan ET, Pena-Castillo L, Chaudhry S, Talukder S, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature Biotechnology. 2009; 27(7):667–70. https://doi.org/10.1038/nbt.1550 PMID: 19561594
10. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499(7457):172–7. https://doi.org/10.1038/nature12311 PMID: 23846655
11. Jolma A, Zhang J, Mondragon E, Morgunova E, Kivioja T, Laverty KU, et al. Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. Genome Res. 2020; 30(7):962–73. https://doi.org/10.1101/gr.258848.119 PMID: 32703884
12. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Molecular Cell. 2014; 54(5):887–900. https://doi.org/10.1016/j.molcel.2014.04.016 PMID: 24837674
13. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nature Methods. 2016; 13(6):508–14. https://doi.org/10.1038/nmeth.3810 PMID: 27018577
14. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010; 141(1):129–41. https://doi.org/10.1016/j.cell.2010.03.009 PMID: 20371350
15. Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol. 2010; 17(7):909–15. https://doi.org/10.1038/nsmb.1838 PMID: 20601959
16. Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biology. 2014; 15(1):R17. https://doi.org/10.1186/gb-2014-15-1-r17 PMID: 24451197
17. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins. PLoS Computational Biology. 2010; 6(7):e1000832–10. https://doi.org/10.1371/journal.pcbi.1000832 PMID: 20617199
18. Hiller M, Pudimat R, Busch A, Backofen R. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. Nucleic Acids Res. 2006; 34(17):e117. https://doi.org/10.1093/nar/gkl544 PMID: 16987907
19. Bahrami-Samani E, Penalva LO, Smith AD, Uren PJ. Leveraging cross-link modification events in CLIP-seq for motif discovery. Nucleic Acids Res. 2015; 43(1):95–103. https://doi.org/10.1093/nar/gku1288 PMID: 25505146
20. Orenstein Y, Wang Y, Berger B. RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data. Bioinformatics. 2016; 32(12):i351–i9. https://doi.org/10.1093/bioinformatics/btw259 PMID: 27307637

21. Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, et al. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Research. 2016; 44(4):e32–e. https://doi.org/10.1093/nar/gkv1025 PMID: 26467480

22. Polishchuk M, Paz I, Yakhini Z, Mandel-Gutfreund Y. SMARTIV: combined sequence and structure denovo motif discovery for in-vivo RNA binding data. Nucleic Acids Res. 2018; 46(W1):W221–W8. https://doi.org/10.1093/nar/gky453 PMID: 29800452

23. Pan X, Shen H-B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinformatics. 2017; 18(1):136. https://doi.org/10.1186/s12859-017-1561-8 PMID: 28245811

24. Heller D, Krestel R, Ohler U, Vingron M, Marsico A. ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data. Nucleic Acids Research. 2017; 45(19):11004–18. https://doi.org/10.1093/nar/gkx756 PMID: 28977546

25. Pietrosanto M, Adinolfi M, Casula R, Ausiello G, Ferre F, Helmer-Citterich M. BEAM web server: a tool for structural RNA motif discovery. Bioinformatics. 2018; 34(6):1058–60. https://doi.org/10.1093/bioinformatics/btx704 PMID: 29095974

26. Munteanu A, Mukherjee N, Ohler U. SSMART: sequence-structure motif identification for RNA-binding proteins. Bioinformatics. 2018; 34(23):3990–8. https://doi.org/10.1093/bioinformatics/bty404 PMID: 29893814

27. Dotu I, Adamson SI, Coleman B, Fournier C, Ricart-Altimiras E, Eyras E, et al. SARNAclust: Semi-automatic detection of RNA protein binding motifs from immunoprecipitation data. PLoS Computational Biology. 2018; 14(3):e1006078–25. https://doi.org/10.1371/journal.pcbi.1006078 PMID: 29596423

28. Ben-Bassat I, Chor B, Orenstein Y. A deep neural network approach for learning intrinsic protein-RNA binding preferences. Bioinformatics. 2018; 34(17):i638–i46. https://doi.org/10.1093/bioinformatics/bty600 PMID: 30423078

29. Zhang C, Lee KY, Swanson MS, Darnell RB. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. Nucleic Acids Res. 2013; 41(14):6793–807. https://doi.org/10.1093/nar/gkt421 PMID: 23685613

30. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA. 2010; 16(6):1096–107. https://doi.org/10.1261/rna.2017210 PMID: 20418358

31. Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(39):14885–90. https://doi.org/10.1073/pnas.0803169105 PMID: 18815376

32. Su Y, Luo Y, Zhao X, Liu Y, Peng J. Integrating thermodynamic and sequence contexts improves protein-RNA binding prediction. PLoS Comput Biol. 2019; 15(9):e1007283. https://doi.org/10.1371/journal.pcbi.1007283 PMID: 31483777

33. Sun L, Xu K, Huang W, Yang YT, Li P, Tang L, et al. Predicting dynamic cellular protein-RNA interactions by deep learning using in vivo RNA structures. Cell Res. 2021; 31(5):495–516. https://doi.org/10.1038/s41422-021-00476-y PMID: 33623109

34. Yan Z, Hamilton WL, Blanchette M. Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. Bioinformatics. 2020; 36(Suppl_1):i276–i84. https://doi.org/10.1093/bioinformatics/btaa456 PMID: 32657407

35. Pelossof R, Singh I, Yang JL, Weirauch MT, Hughes TR, Leslie CS. Affinity regression predicts the recognition code of nucleic acid&ndash;binding proteins. Nature Biotechnology. 2015; 33(12):1242–9. https://doi.org/10.1038/nbt.3343 PMID: 26571099

36. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. Annu Rev Biochem. 2010; 79:233–69. https://doi.org/10.1146/annurev-biochem-060408-091030 PMID: 20334529

37. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. Nature. 2009; 461(7268):1248–53. https://doi.org/10.1038/nature08473 PMID: 19865164

38. Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nature Reviews Molecular Cell Biology. 2018; 19(5):327–41. https://doi.org/10.1038/nrm.2017.130 PMID: 29339797

39. Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, et al. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Molecular Cell. 2018; 70(5):854–67.e9. https://doi.org/10.1016/j.molcel.2018.05.001 PMID: 29883606

40. Sanchez de Groot N, Armaos A, Grana-Montes R, Alriquet M, Calloni G, Vabulas RM, et al. RNA structure drives interaction with proteins. Nat Commun. 2019; 10(1):3246. https://doi.org/10.1038/s41467-019-10923-5 PMID: 31324771

41. Lewis CJ, Pan T, Kalsotra A. RNA modifications and structures cooperate to guide RNA-protein interactions. Nat Rev Mol Cell Biol. 2017; 18(3):202–10. https://doi.org/10.1038/nrm.2016.163 PMID: 28144031

42. Corley M, Burns MC, Yeo GW. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Mol Cell. 2020; 78(1):9–29. https://doi.org/10.1016/j.molcel.2020.03.011 PMID: 32243832

43. Bernhart SH, Hofacker IL, Stadler PF. Local RNA base pairing probabilities in large sequences. Bioinformatics. 2006; 22(5):614–5. https://doi.org/10.1093/bioinformatics/btk014 PMID: 16368769

44. Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA. 2008; 14(6):1164–73. https://doi.org/10.1261/rna.894608 PMID: 18456842

45. Wilson D, Martinez T. Improved heterogeneous distance functions. J Artif Intell Res. 1997; 6:1–34.

46. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods. 2012; 9(5):471–2. https://doi.org/10.1038/nmeth.1938 PMID: 22426491

47. Masliah G, Barraud P, Allain FH. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cell Mol Life Sci. 2013; 70(11):1875–95. https://doi.org/10.1007/s00018-012-1119-x PMID: 22918483

48. Corley M, Solem A, Qu K, Chang HY, Laederach A. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. Nucleic Acids Research. 2015; 43(3):1859–68. https://doi.org/10.1093/nar/gkv010 PMID: 25618847

49. Dimitrova-Paternoga L, Jagtap PKA, Chen PC, Hennig J. Integrative Structural Biology of Protein-RNA Complexes. Structure. 2020; 28(1):6–28. https://doi.org/10.1016/j.str.2019.11.017 PMID: 31864810

50. Batey RT, Rambo RP, Doudna JA. Tertiary Motifs in RNA Structure and Folding. Angew Chem Int Ed Engl. 1999; 38(16):2326–43. https://doi.org/10.1002/(sici)1521-3773(19990816)38:16<2326::aid-anie2326>3.0.co;2-3 PMID: 10458781

51. Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM. The roles of structural dynamics in the cellular functions of RNAs. Nat Rev Mol Cell Biol. 2019; 20(8):474–89. https://doi.org/10.1038/s41580-019-0136-0 PMID: 31182864

52. Ustianenko D, Chiu HS, Treiber T, Weyn-Vanhentenryck SM, Treiber N, Meister G, et al. LIN28 Selectively Modulates a Subclass of Let-7 MicroRNAs. Mol Cell. 2018; 71(2):271–83 e5. https://doi.org/10.1016/j.molcel.2018.06.029 PMID: 30029005

53. Wilbert ML, Huelga SC, Kapeli K, Stark TJ, Liang TY, Chen SX, et al. LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Mol Cell. 2012; 48(2):195–206. https://doi.org/10.1016/j.molcel.2012.08.004 PMID: 22959275

54. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc. 2017; 12(2):255–78. https://doi.org/10.1038/nprot.2016.169 PMID: 28079879

55. Samsonova A, El Hage K, Desforges B, Joshi V, Clement MJ, Lambert G, et al. Lin28, a major translation reprogramming factor, gains access to YB-1-packaged mRNA through its cold-shock domain. Commun Biol. 2021; 4(1):359. https://doi.org/10.1038/s42003-021-01862-3 PMID: 33742080

56. Zhang J, Fan JS, Li S, Yang Y, Sun P, Zhu Q, et al. Structural basis of DNA binding to human YB-1 cold shock domain regulated by phosphorylation. Nucleic Acids Res. 2020; 48(16):9361–71. https://doi.org/10.1093/nar/gkaa619 PMID: 32710623

57. Mayr F, Schutz A, Doge N, Heinemann U. The Lin28 cold-shock domain remodels pre-let-7 microRNA. Nucleic Acids Res. 2012; 40(15):7492–506. https://doi.org/10.1093/nar/gks355 PMID: 22570413

58. Siebert M, Soding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. Nucleic Acids Res. 2016; 44(13):6055–69. https://doi.org/10.1093/nar/gkw521 PMID: 27288444

59. Wong KC, Chan TM, Peng C, Li Y, Zhang Z. DNA motif elucidation using belief propagation. Nucleic Acids Res. 2013; 41(16):e153. https://doi.org/10.1093/nar/gkt574 PMID: 23814189

60. Pietrosanto M, Mattei E, Helmer-Citterich M, Ferre F. A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications. Nucleic Acids Res. 2016; 44(18):8600–9. https://doi.org/10.1093/nar/gkw750 PMID: 27580722

61. Li B, Cao Y, Westhof E, Miao Z. Advances in RNA 3D Structure Modeling Using Experimental Data. Front Genet. 2020; 11:574485. https://doi.org/10.3389/fgene.2020.574485 PMID: 33193680

62. Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. RNA. 2001; 7(4):499–512. https://doi.org/10.1017/s1355838201002515 PMID: 11345429

63. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun. 2021; 12(1):941. https://doi.org/10.1038/s41467-021-21194-4 PMID: 33574226

64. Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. 2020; 30(2):214–26. https://doi.org/10.1101/gr.247494.118 PMID: 31992613

65. Calabrese C, Davidson NR, Iu DDxF, Fonseca NA, He Y, Kahles AxE, et al. Genomic basis for RNA alterations in cancer. Nature. 2020:1–50. https://doi.org/10.1038/s41586-020-1970-0 PMID: 32025019

66. Sundararajan M, Taly A, Yan Q, editors. Axiomatic Attribution for Deep Networks. Proceedings of the 34th International Conference on Machine Learning, PMLR; 2017; Sydney, NSW, Australia.

67. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, et al. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 2013; 41(Database issue):D110–7. https://doi.org/10.1093/nar/gks1058 PMID: 23161672

68. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

69. Lange SJ, Maticzka D, Mohl M, Gagnon JN, Brown CM, Backofen R. Global or local? Predicting secondary structure and accessibility in mRNAs. Nucleic Acids Res. 2012; 40(12):5215–26. https://doi.org/10.1093/nar/gks181 PMID: 22373926

70. Carvalho LE, Lawrence CE. Centroid estimation in discrete high-dimensional spaces with applications in biology. Proc Natl Acad Sci U S A. 2008; 105(9):3209–14. https://doi.org/10.1073/pnas.0712329105 PMID: 18305160

71. Floyd R. Algorithm-97—Shortest Path. Communications of ACM. 1962; 5(6):345.

72. Warshall S. A Theorem on Boolean Matrices. Journal of Acm. 1962; 9(1).

73. Stanfill C, Waltz D. Toward memory-based reasoning. Communication of ACM. 1986; 29(12).

74. Schones DE, Sumazin P, Zhang MQ. Similarity of position frequency matrices for transcription factor binding sites. Bioinformatics. 2005; 21(3):307–13. https://doi.org/10.1093/bioinformatics/bth480 PMID: 15319260

75. Bron C, Kerbosch j. Finding All Cliques of an Undirected Graph. Communications of ACM 1973; 16(9).

76. Calzals F, Karande C. A note on the problem of reporting maximal cliques.. Theor Comput Sci 2008;407.

77. Tomita E TA, Takahashi H.. The worst-case time complexity for generating all maximal cliques and computational experiments.. Theor Comput Sci 2006; 363((1):28–42.

78. Ding F, Lavender CA, Weeks KM, Dokholyan NV. Three-dimensional RNA structure refinement by hydroxyl radical probing. Nature Methods. 2012; 9(6):603–8. https://doi.org/10.1038/nmeth.1976 PMID: 22504587

79. Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, et al. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Research. 2016; 44(7): e63–e. https://doi.org/10.1093/nar/gkv1479 PMID: 26687716

80. Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, et al. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. J Chem Theory Comput. 2016; 12(1):405–13. https://doi.org/10.1021/acs.jctc.5b00935 PMID: 26631602

81. Hoover WG. Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A Gen Phys. 1985; 31(3):1695–7. https://doi.org/10.1103/physreva.31.1695 PMID: 9895674

82. Martonak R, Molteni C, Parrinello M. Ab initio molecular dynamics with a classical pressure reservoir: simulation of pressure-induced amorphization in a Si35H36 cluster. Phys Rev Lett. 2000; 84(4):682–5. https://doi.org/10.1103/PhysRevLett.84.682 PMID: 11017346

83. Darden T, Perera L, Li L, Pedersen L. New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. Structure. 1999; 7(3):R55–60. https://doi.org/10.1016/s0969-2126(99)80033-1 PMID: 10368306

84. Hess B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. J Chem Theory Comput. 2008; 4(1):116–22. https://doi.org/10.1021/ct700200b PMID: 26619985

85. Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013; 29(7):845–54. https://doi.org/10.1093/bioinformatics/btt055 PMID: 23407358

86. Klauda JB, Venable RM, Freites JA, O'Connor JW, Tobias DJ, Mondragon-Ramirez C, et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. J Phys Chem B. 2010; 114(23):7830–43. https://doi.org/10.1021/jp101759q PMID: 20496934

87. Huang J, MacKerell AD Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. J Comput Chem. 2013; 34(25):2135–45. https://doi.org/10.1002/jcc.23354 PMID: 23832629

88. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods. 2017; 14(1):71–3. https://doi.org/10.1038/nmeth.4067 PMID: 27819658

89. Jorgensen WL, Chandrasekhar J, Buckner JK, Madura JD. Computer simulations of organic reactions in solution. Ann N Y Acad Sci. 1986; 482:198–209. https://doi.org/10.1111/j.1749-6632.1986.tb20951.x PMID: 3471104

90. Vajda S, Yueh C, Beglov D, Bohnuud T, Mottarella SE, Xia B, et al. New additions to the ClusPro server motivated by CAPRI. Proteins. 2017; 85(3):435–44. https://doi.org/10.1002/prot.25219 PMID: 27936493

91. Sun L, Fazal FM, Li P, Broughton JP, Lee B, Tang L, et al. RNA structure maps across mammalian cellular compartments. Nat Struct Mol Biol. 2019; 26(4):322–30. https://doi.org/10.1038/s41594-019-0200-7 PMID: 30886404