RNA NetMotif: identifying sequence-structure RNA network motifs in RNA-protein binding sites

Hongli Ma^{1,2,3,5,†}, Han Wen^{6,†}, Zhiyuan Xue⁷, Guojun Li^{1,5,8,*}, Zhaolei Zhang^{2,3,4,*}

1. Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, 266237, China
 2. Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada
 3. Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada
 4. Department of Computer Science, University of Toronto, Toronto, ON M5S 3E1, Canada
 5. School of Mathematics, Shandong University, Jinan, 250100, China
 6. DP Technology, Beijing, 100080
 7. West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, 610041, China
 8. School of Mathematical Science, Liaocheng University, Liaocheng, 252000, China
- [†] These authors contributed equally to this work.

*** Corresponding authors**

Zhaolei Zhang
Zhaolei.Zhang@utoronto.ca

Guojun Li:
guojunsdu@gmail.com

Abstract

RNA molecules can adopt stable secondary and tertiary structures, which is 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 discrete molecular dynamics folding simulations for selected 4 RBPs, and RNA-protein docking for LIN28. The simulation results, e.g., solvent accessibility and energetics, further support the biological relevance of the discovered network subgraphs.

Keywords

RNA binding protein (RBP), graphs, RNA structure

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 specificity to specific RBPs. It is known that RNAs form energetically favorable secondary structures, which can be represented as a graph, 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 Subgraphs) and GraphK graph search 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 regulates every aspect of the RNA biogenesis and RNA biology, including RNA splicing, modification, degradation, protein translation, and RNA subcellular localization (1-6). 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 (7-9). 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 (10-12), and in vivo methods such as PAR-CLIP, iCLIP, and eCLIP can identify regions on the RNA transcript that are bound by RBPs in cells (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 had 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 (16-33). A number of these representative methods are briefly described below.

The 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 score 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 captured structure inter-dependencies between neighboring nucleotides through graph based encoding or hidden Markov model. Moreover, methods such as BEAM (25) attempted to explicitly encode the pattern of nucleotide base pairing and catalogue the occurrence of specific secondary structure motifs based on the length of the stem or the size of the hairpin loop. SARNAClust (27) comprehensively studied the topology and annotations of the complete RNA structure and provides several options for graph transformations of RNA shapes. 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 alone. 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 (34,35). 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 2-D or 3-D structure motifs that are essential in the stabilization of the RNA molecule and presenting the nucleotide moieties to RBPs (36-39). The 2-D 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 3-D 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 2-D sequence entities as RNA network motifs.

Interactions between RNA-binding proteins (RBPs) and RNA molecules employ diverse and dynamic modes (40), 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. Second, given that RNA structures are represented as 2-D or 3-D 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. (41). (ii) We developed a novel GraphK algorithm, which can break apart 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 discrete molecular dynamics (MD) folding simulation and evaluated the significance of the discovered network motifs by comparing their spatial exposure with other regions (42).

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 the 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.ccb.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. MATERIALS AND METHODS

2.1 Collection of RBP eCLIP data and preprocessing

We downloaded the “enhanced UV crosslinking followed by immunoprecipitation” (eCLIP) data (13) 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 (1).

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 rationale 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) (43), 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 1**. 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 (44,45) at 80% similarity cut-off.

2.2 Overall workflow of the RNANetMotif method

As shown in **Figure 1**, 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 filter to obtain the final EKS pool. In Step 3, we compute the HVDM (Heterogeneous Value Difference Metric) distance matrix (46) 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.

2.3 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 (41). 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 (47). 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 (48).

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 (W=100, L=100, u=1) were used. 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.

2.4 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 (**Figure 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 **Figure 2**, the linear sequence of the local folded RNA follows the following direction: $5' \rightarrow s \rightarrow s' \rightarrow t' \rightarrow t \rightarrow 3'$. 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 subgraph 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 $O(|U|^3)$ (49,50). 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 EKSSes into three categories: opposite direction extensions, mixed-mode extensions, and same direction extensions. As shown at the top of **Figure 2**, the opposite direction extensions can generate EKSSes that are either left-opened or right-opened subgraphs. Mixed-mode extensions have the most diverse shapes and topology. 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 EKSSes have lower chance of forming base-pairing unless pseudoknot structure is allowed.

2.5 GraphK – a novel subgraph extraction approach

Having defined EKS, we next introduce an algorithm, GraphK, which applies extension, extraction, and filtering steps to break a complex graph into overlapping representative EKSSes. The detailed steps of the GraphK algorithm are as follows (**Figure 1 (Step 2)**). In Step 1, we conduct k -mer extension as described above in each RBP-bound RNA sequence to obtain all EKSSes associated with the RNA sequence. In addition, we added an annotation for each EKS, including its type, compactness score, and the starting positions of the two k -mers. In **Step 2**, we extract EKSSes generated from opposite direction extensions (i.e., right-opened or left-opened), and exclude those generated by opposite direction and mixed mode extensions. The rationale is to extract only those likely to form single stranded RNAs (see above). In **Step 3**, we filter EKSSes by their compactness scores and only retain those with lower compactness scores, i.e., those

with less base pairings. We next filtered the EKSeS by the position of the nucleotides and kept the EKSeS whose sequence is entirely located in 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 2-D or 3-D 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 (11), 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 **Figure 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.

2.6 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 (46,51). Given two EKSeS with the feature vector $\vec{x} = (x_i, i = 1, 2, \dots, 12k)$ and $\vec{y} = (y_i, i = 1, 2, \dots, 12k)$, 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: $classes = \{C_i, i = 1, 2, 3, \dots, 10\}$. As shown in **Figure 1 (Step 3)**, 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 $\vec{x}$ and $\vec{y}$ is defined in Equation (1), which considers the correlations between each possible attribute value and each EKS class.

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

where

$$N_{\alpha, x_{\alpha}} = \sum_{h=1}^{\#classes} N_{\alpha, x_{\alpha}, C_h} \quad (2)$$

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

$N_{\alpha, x_{\alpha}, C_h}$ is the number of instances in all EKSeS that have value x_{α} for attribute α and class C_h ; $N_{\alpha, x_{\alpha}}$ 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:

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

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

$$HVDM(\vec{x}, \vec{y}) = \sqrt{\sum_{a=1}^{12k} d_a^2(x_a, y_a)} \quad (5)$$

where

$$d_a(x_a, y_a) = \begin{cases} vdm_a(\vec{x}, \vec{y}), & \text{if } a \text{ is categorical} \\ diff_a(\vec{x}, \vec{y}), & \text{if } a \text{ is continuous} \\ 1, & \text{other conditions} \end{cases} \quad (6)$$

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 Equation (7). N_{v_1} and N_{v_2} are the neighbor sets of v_1 and v_2 respectively.

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

In the second pruning step, we removed any edges e with weight lower than a cutoff, $w_e \leq w_{cutoff}$; the purpose of this was to retain only those sufficiently similar EKS candidates. After experimenting with different parameters, we set $n = 2\% \times \text{size of } V_S$, $w_{cutoff} = 0.9 \text{ quantile of all weights}$. We denote the pruned graph as the similarity network among filtered EKSES. In **Supplementary Table 1**, we summarize the number of the nodes and edges of the similarity network for the 22 RBPs.

2.7 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 (52) 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 **Supplementary Method S1** 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 (53). 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 (54–56) 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.

2.8 RNA 3D structure modeling and evaluation of network motifs

The RNA 3D structures were predicted using a discrete molecular dynamics (DMD) method (42) with predicted base-pair information provided as constraints. The DMD engine and scripts were adopted from the literature (57) and kindly provided by Dr. Feng Ding's lab (<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 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 (58) was performed on such neighboring counts of identified motifs against the average of all the k-mer regions.

Charmm-gui webserver (59-61) was used to set up the MD systems, the complex was solvated in a water box with 15 Å buffer of water extending from the RNA or protein-RNA complex. K⁺ and Cl²⁻ 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 (62,63) was used with temperature T = 30°C. The Parrinello-Rahman method (64) was used for pressure coupling. A 10-Å switching distance and a 12-Å cutoff distance were used for non-bonded interactions. The particle mesh Ewald method (65) was used for electrostatics calculations. The LINCS algorithm (66) was used to constrain the hydrogen containing bond lengths for a 2-fs time step. The energy minimization and MD simulations were performed with the GROMACS program (67) version 2019.6-CPU/GPU using the CHARMM36m force field (68-70) and TIP3P water model (71).

The RNA structures after 10-ns production run were used as receptors and the LIN28B CSD structure (PDB ID 4A4I) (72) was used as ligand to perform protein-RNA docking using the ClusPro webserver (73-75); 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 (**Figure S2**). 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 $<3.5 \text{ \AA}$, and the deviation of the donor-hydrogen-acceptor angle from 180° is $<60^\circ$. 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 (76).

3. RESULTS

3.1 Selection and optimization of k in GraphK partition

As shown in **Figure 1 (Step 2)** and **Figure 2**, the EKSeS were constructed by traversing and extending k nucleotides from a specific base-pair. 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 **Figure 3 (top panel)**, approximately 79.51% of all the gap lengths is 1, which represents stacked adjacent base pairs. The bottom of **Figure 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 EKSe. 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 **Figure 3** shows that the number of occurrences decreases as the gap length increases.

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: P (paired), H (hairpin Loop), E (extended, unstructured), I (internal loop, bulge), M (multiple loop). We next counted the occurrence of sequence and structure of EKSeS respectively; **Figure 4** shows the frequency of top sequence instances and top secondary structure instances in scatter plots for five selected RBPs on different k . From **Figure 4**, we find that after applying GraphK partition, the distribution of top sequence instances of EKSe has lower frequencies 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 (**Method 2.6, 2.7**).

3.2 Significant modules in the similarity network of EKSeS

We next calculated HVDM distances among pairs of EKSeS and constructed a similarity network with individual EKS as nodes and HVDM distance as edge weight. 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. As described above in section 3.1, we next discretized secondary structure representation of each nucleotide and visualized the sequence-structure preference of these significant modules in **Figure 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 RBPs than single-stranded RBPs, which is consistent with previous observations (40,77). We also did not observe significant EKS modules for two other RBPs PPIG, BCLAF1, which may be because these RBPs have multiple binding modes or multiple RNA binding domains, generating a mixture of signals (78). Further investigation is warranted for these RBPs.

Figure 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 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.

3.3 Significance of EKS motifs tested by 3D structure modelling

In the world of protein, linear polypeptides usually fold into certain three-dimensional structures to ultimately achieve cellular functions. Similarly, protein-RNA interactions often require RNA to fold into specific 3-D shapes in order to precisely spatially present relevant chemical moieties on sugar-phosphate backbone or nucleoside bases (79). 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 an important role in interaction between RNA and other molecules (80). Towards this goal, we performed large scale discrete molecular dynamics simulations to model the 3-D structure of four well-studied RNA binding proteins (G3BP1, RBM15, LIN28B, 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 2.8), 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 four well-studied RBPs (G3BP1, LIN28B, PUM2, RBM15), we performed one-sided Mann Whitney U Test (Wilcoxon Rank Sum Test) to compare the atom counts of the identified motif with all other k-mer regions. **Figure 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.

3.4 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 (81,82).

We chose two representative sequences from the 5mer motifs and one from 4mer motifs to perform all-atom structure refinement and docking (see **Methods**). We chose the ClusPro web server (74) 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. 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 **Figure 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) (**Supplementary Table 2, Figure 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 mutation studies indicated that K88, K89 (K98, K99 in LIN28A) play a role in electrostatic interactions with nucleic acids (83), and a recent study found a highly conserved Lys residue in YB1 protein (equivalent to K92 in Lin28B) is involved in ssDNA binding (84), as did the structural study of LIN28 on ssRNA nucleotides (72). Altogether, these evidences 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.

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

We next compared RNANetMotif with three other computational methods, BEAM, GraphProt and ssHMM with two different secondary structure predictors on the eCLIP datasets. We displayed the motifs identified by RNANetMotif and by other methods in **Figure 8** and summarized the common motif elements in the right column. As seen in **Figure 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.

3.6 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 2**, 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 3**, 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 **Supplementary Materials**.

3.7 Run-time and memory comparisons

We compared the run time of RNANetMotif with other methods on five eCLIP datasets on a 32-core CPU (see Supplemental **Figure S3**). We excluded the time for RNA secondary structure prediction and model training time for Graphprot. RNANetMotif consumes 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 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 regard to this task.

4. 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 3-dimensional structures of a representative set of RBP-RNA complexes and derive a set of common 3-D 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 (85,86). A number of methods such as BEAM or GraphProt have improved upon PWMs by adding structural descriptors to each position, i.e., helix, stem, loop, etc, which has showed improvement (16,25). 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 bond and other interactions among amino acid residues.

We introduced a new concept to represent RNA elements, i.e., EKS (Extended K-mer Subgraphs) and an algorithm GraphK. As demonstrated in this study, we think EKS is a promising approach in mining RNA secondary structures for enriched motifs. We further showed in **Result 3.6** 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 with other methods, RNANetMotif has the following unique characteristics. (i) It recognizes RNA target sequences as a network 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. The 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.ccb.utoronto.ca>.

There are several limitations and avenues for future improvement in 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 (87,88). This is mostly due to the lack of high-resolution 3-D RNA structures, which makes accurate prediction of these non-canonical interactions less feasible. We anticipate deep learning models, especially graph models, can probably alleviate this challenge in the coming years(33,89-91). 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.

5. DATA AVAILABILITY

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.ccb.utoronto.ca>).

6. 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 Council of Canada (Discovery Grant, RGPIN-2017-06743)

7. ACKNOWLEDGEMENTS

This research was enabled in part by support provided by Compute Canada (www.computecanada.ca). Computations 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. HM thanks Shandong University for their financial support through scholarships under overseas exchange program. HW would like to thank Dr. Philip Kim from CCBR, University of Toronto for post-doctoral fellowship support in the early stage of this work. We thank Dr. Feng Ding and Zhenzhen Zhang from Clemson University for their support and useful discussion about DMD method.

8. AUTHOR CONTRIBUTIONS

HM and ZZ conceived the project. HM collected the data and developed the algorithm. HW conducted molecular dynamics and modeling. ZX implemented the webserver. HM, HW and ZZ wrote the paper. GL and ZZ supervised the project.

Tables

Table 1. RBPs and the dataset 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	WD1, WD2, WD3, WD4, WD5, WD6, WD7	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, Pumilio 1, Pumilio 2, Pumilio 3, Pumilio 4, Pumilio 5, Pumilio 6, Pumilio 7, Pumilio 8	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

Table 2. Implanted Motifs.

Implanted Motifs		
CCACCA in hairpin loop	NNNCCACCANN	Sequence
	(((((.....)))	Structure
AUG in bulge	NNNAUGNNNNNNNNNNNNNNNNNN	Sequence
	(((((...(((.....))))))))))	Structure
AA_AA in internal loop	NNNAANNNNNNNAANN	Sequence
	(((((...((()))...)))	Structure
GAGAGAGA in hairpin loop	NNNNNGAGAGAGANNNN	Sequence
	(((((.....))))))	Structure
CCCC_AAAA_CCCC in multiloop	NNNCCCCNNNNNNNAAAANNNNNNCC CCNN	Sequence
	(((((...((()))...))) ((((((()))...)))	Structure

Table 3. Recovery rate of sequence and structure on synthetic data for four methods

Software (type)	3nt bulge loop	4nt internal loop	6nt hairpin loop	8nt hairpin loop	12nt multi loop	Overall recover rate
RNASeqMotif (Sequence)	1	1	1	1	1	1
RNASeqMotif (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

References

1. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.Y., Cody, N.A.L., Dominguez, D. *et al.* (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**, 711-719.
2. Neelamraju, Y., Hashemikhabir, S. and Janga, S.C. (2015) The human RBPome: from genes and proteins to human disease. *J Proteomics*, **127**, 61-70.
3. Matia-Gonzalez, A.M., Laing, E.E. and Gerber, A.P. (2015) Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat Struct Mol Biol*, **22**, 1027-1033.
4. Gerstberger, S., Hafner, M. and Tuschl, T. (2014) A census of human RNA-binding proteins. *Nat Rev Genet*, **15**, 829-845.
5. Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393-1406.
6. Baltz, A.G., Munschauer, M., Schwanhauser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*, **46**, 674-690.
7. Qin, H., Ni, H., Liu, Y., Yuan, Y., Xi, T., Li, X. and Zheng, L. (2020) RNA-binding proteins in tumor progression. *J Hematol Oncol*, **13**, 90.
8. Gebauer, F., Schwarzl, T., Valcarcel, J. and Hentze, M.W. (2021) RNA-binding proteins in human genetic disease. *Nat Rev Genet*, **22**, 185-198.
9. Kechavarzi, B. and Janga, S.C. (2014) Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol*, **15**, R14.
10. Ray, D., Kazan, H., Chan, E.T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q. and Hughes, T.R. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*, **27**, 667-670.
11. Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A. and Burge, C.B. (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell*, **54**, 887-900.
12. Jolma, A., Zhang, J., Mondragon, E., Morgunova, E., Kivioja, T., Laverty, K.U., Yin, Y., Zhu, F., Bourenkov, G., Morris, Q. *et al.* (2020) Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res*, **30**, 962-973.
13. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*, **13**, 508-514.
14. Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M. and Ule, J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, **17**, 909-915.
15. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129-141.
16. Maticzka, D., Lange, S.J., Costa, F. and Backofen, R. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*, **15**, R17.

17. Kazan, H., Ray, D., Chan, E.T., Hughes, T.R. and Morris, Q. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*, **6**, e1000832.
18. Hiller, M., Pudimat, R., Busch, A. and Backofen, R. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res*, **34**, e117.
19. Bahrami-Samani, E., Penalva, L.O., Smith, A.D. and Uren, P.J. (2015) Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Res*, **43**, 95-103.
20. Orenstein, Y., Wang, Y. and Berger, B. (2016) RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data. *Bioinformatics*, **32**, i351-i359.
21. Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C. and Zeng, J. (2016) A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res*, **44**, e32.
22. Polishchuk, M., Paz, I., Yakhini, Z. and Mandel-Gutfreund, Y. (2018) SMARTIV: combined sequence and structure de-novo motif discovery for in-vivo RNA binding data. *Nucleic Acids Res*, **46**, W221-W228.
23. Pan, X. and Shen, H.B. (2017) RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics*, **18**, 136.
24. Heller, D., Krestel, R., Ohler, U., Vingron, M. and Marsico, A. (2017) ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data. *Nucleic Acids Res*, **45**, 11004-11018.
25. Pietrosanto, M., Adinolfi, M., Casula, R., Ausiello, G., Ferre, F. and Helmer-Citterich, M. (2018) BEAM web server: a tool for structural RNA motif discovery. *Bioinformatics*, **34**, 1058-1060.
26. Munteanu, A., Mukherjee, N. and Ohler, U. (2018) SSMART: sequence-structure motif identification for RNA-binding proteins. *Bioinformatics*, **34**, 3990-3998.
27. Dotu, I., Adamson, S.I., Coleman, B., Fournier, C., Ricart-Altimiras, E., Eyra, E. and Chuang, J.H. (2018) SARNACLUST: Semi-automatic detection of RNA protein binding motifs from immunoprecipitation data. *PLoS Comput Biol*, **14**, e1006078.
28. Ben-Bassat, I., Chor, B. and Orenstein, Y. (2018) A deep neural network approach for learning intrinsic protein-RNA binding preferences. *Bioinformatics*, **34**, i638-i646.
29. Zhang, C., Lee, K.Y., Swanson, M.S. and Darnell, R.B. (2013) Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res*, **41**, 6793-6807.
30. Li, X., Quon, G., Lipshitz, H.D. and Morris, Q. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096-1107.
31. Rabani, M., Kertesz, M. and Segal, E. (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci U S A*, **105**, 14885-14890.
32. Su, Y., Luo, Y., Zhao, X., Liu, Y. and Peng, J. (2019) Integrating thermodynamic and sequence contexts improves protein-RNA binding prediction. *PLoS Comput Biol*, **15**, e1007283.
33. Sun, L., Xu, K., Huang, W., Yang, Y.T., Li, P., Tang, L., Xiong, T. and Zhang, Q.C. (2021) Predicting dynamic cellular protein-RNA interactions by deep learning using in vivo RNA structures. *Cell Res*, **31**, 495-516.
34. Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248-1253.

35. Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B. and Mann, R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*, **79**, 233-269.
36. Hentze, M.W., Castello, A., Schwarzl, T. and Preiss, T. (2018) A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol*, **19**, 327-341.
37. Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A. *et al.* (2018) Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell*, **70**, 854-867 e859.
38. Sanchez de Groot, N., Armaos, A., Grana-Montes, R., Alriquet, M., Calloni, G., Vabulas, R.M. and Tartaglia, G.G. (2019) RNA structure drives interaction with proteins. *Nat Commun*, **10**, 3246.
39. Lewis, C.J., Pan, T. and Kalsotra, A. (2017) RNA modifications and structures cooperate to guide RNA-protein interactions. *Nat Rev Mol Cell Biol*, **18**, 202-210.
40. Corley, M., Burns, M.C. and Yeo, G.W. (2020) How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol Cell*, **78**, 9-29.
41. Bernhart, S.H., Hofacker, I.L. and Stadler, P.F. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614-615.
42. Ding, F., Sharma, S., Chalasani, P., Demidov, V.V., Broude, N.E. and Dokholyan, N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164-1173.
43. Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.J., Lopez, G., Valencia, A. and Tress, M.L. (2013) APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res*, **41**, D110-117.
44. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-1659.
45. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152.
46. Wilson, D.R. and Martinez, T.R. (1997) Improved heterogeneous distance functions. *J Artif Intell Res*, **6**, 1-34.
47. Lange, S.J., Maticzka, D., Mohl, M., Gagnon, J.N., Brown, C.M. and Backofen, R. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res*, **40**, 5215-5226.
48. Carvalho, L.E. and Lawrence, C.E. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc Natl Acad Sci U S A*, **105**, 3209-3214.
49. Floyd, R.W. (1962) Algorithm-97 - Shortest Path. *Communications of the Acm*, **5**, 345-345.
50. Warshall, S. (1962) A Theorem on Boolean Matrices. *J Acm*, **9**, 11-&.
51. Stanfill, C. and Waltz, D. (1986) Toward memory-based reasoning. *Communications of the ACM*, **29**, 1213-1228.
52. Nepusz, T., Yu, H. and Paccanaro, A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods*, **9**, 471-472.
53. Schones, D.E., Sumazin, P. and Zhang, M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307-313.
54. Bron, C. and Kerbosch, J. (1973) Finding All Cliques of an Undirected Graph [H]. *Communications of the Acm*, **16**, 575-577.
55. Calzals, F. and Karande, C. (2008) A note on the problem of reporting maximal cliques. *Theor Comput Sci*, **407**, 564-568.
56. Tomita, E., Tanaka, A. and Takahashi, H. (2006) The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor Comput Sci*, **363**, 28-42.

57. Ding, F., Lavender, C.A., Weeks, K.M. and Dokholyan, N.V. (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nature Methods*, **9**, 603-608.
58. Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, **1**, 80-83.
59. Brooks, B.R., Brooks III, C.L., Mackerell Jr., A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. *et al.* (2009) CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, **30**, 1545-1614.
60. Jo, S., Kim, T., Iyer, V.G. and Im, W. (2008) CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, **29**, 1859-1865.
61. Lee, J., Cheng, X., Swails, J.M., Yeom, M.S., Eastman, P.K., Lemkul, J.A., Wei, S., Buckner, J., Jeong, J.C., Qi, Y. *et al.* (2016) CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation*, **12**, 405-413.
62. Nosé, S. (1984) A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, **52**, 255-268.
63. Hoover, W.G. (1985) Canonical dynamics: Equilibrium phase-space distributions. *Physical review. A, General physics*, **31**, 1695-1697.
64. Parrinello, M. and Rahman, A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, **52**, 7182-7190.
65. Darden, T., York, D. and Pedersen, L. (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, **98**, 10089-10092.
66. Hess, B., Bekker, H., Berendsen, H.J.C. and Fraaije, J.G.E.M. (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, **18**, 1463-1472.
67. Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D. *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**, 845-854.
68. Klauda, J.B., Venable, R.M., Freites, J.A., O'Connor, J.W., Tobias, D.J., Mondragon-Ramirez, C., Vorobyov, I., MacKerell, A.D., Jr. and Pastor, R.W. (2010) Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *The journal of physical chemistry. B*, **114**, 7830-7843.
69. Huang, J. and MacKerell, A.D., Jr. (2013) CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem*, **34**, 2135-2145.
70. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B.L., Grubmüller, H. and MacKerell, A.D. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, **14**, 71-73.
71. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, **79**, 926-935.
72. Mayr, F., Schutz, A., Doge, N. and Heinemann, U. (2012) The Lin28 cold-shock domain remodels pre-let-7 microRNA. *Nucleic Acids Res*, **40**, 7492-7506.
73. Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R. and Vajda, S. (2013) How good is automated protein docking? *Proteins*, **81**, 2159-2166.
74. Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorney, D., Yueh, C., Beglov, D. and Vajda, S. (2017) The ClusPro web server for protein-protein docking. *Nat Protoc*, **12**, 255-278.
75. Vajda, S., Yueh, C., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R. and Kozakov, D. (2017) New additions to the ClusPro server motivated by CAPRI. *Proteins*, **85**, 435-444.

76. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics. *Journal of molecular graphics*, **14**, 33-38, 27-38.
77. Masliah, G., Barraud, P. and Allain, F.H. (2013) RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell Mol Life Sci*, **70**, 1875-1895.
78. Dimitrova-Paternoga, L., Jagtap, P.K.A., Chen, P.C. and Hennig, J. (2020) Integrative Structural Biology of Protein-RNA Complexes. *Structure*, **28**, 6-28.
79. Batey, R.T., Rambo, R.P. and Doudna, J.A. (1999) Tertiary Motifs in RNA Structure and Folding. *Angew Chem Int Ed Engl*, **38**, 2326-2343.
80. Ganser, L.R., Kelly, M.L., Herschlag, D. and Al-Hashimi, H.M. (2019) The roles of structural dynamics in the cellular functions of RNAs. *Nat Rev Mol Cell Biol*, **20**, 474-489.
81. Ustianenko, D., Chiu, H.S., Treiber, T., Weyn-Vanhentenryck, S.M., Treiber, N., Meister, G., Sumazin, P. and Zhang, C. (2018) LIN28 Selectively Modulates a Subclass of Let-7 MicroRNAs. *Mol Cell*, **71**, 271-283 e275.
82. Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T. *et al.* (2012) LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Mol Cell*, **48**, 195-206.
83. Samsonova, A., El Hage, K., Desforges, B., Joshi, V., Clement, M.J., Lambert, G., Henrie, H., Babault, N., Craveur, P., Maroun, R.C. *et al.* (2021) Lin28, a major translation reprogramming factor, gains access to YB-1-packaged mRNA through its cold-shock domain. *Commun Biol*, **4**, 359.
84. Zhang, J., Fan, J.S., Li, S., Yang, Y., Sun, P., Zhu, Q., Wang, J., Jiang, B., Yang, D. and Liu, M. (2020) Structural basis of DNA binding to human YB-1 cold shock domain regulated by phosphorylation. *Nucleic Acids Res*, **48**, 9361-9371.
85. Siebert, M. and Seding, J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, **44**, 6055-6069.
86. Wong, K.C., Chan, T.M., Peng, C.B., Li, Y. and Zhang, Z.L. (2013) DNA motif elucidation using belief propagation. *Nucleic Acids Research*, **41**.
87. Li, B., Cao, Y., Westhof, E. and Miao, Z. (2020) Advances in RNA 3D Structure Modeling Using Experimental Data. *Front Genet*, **11**, 574485.
88. Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499-512.
89. Ghanbari, M. and Ohler, U. (2020) Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res*, **30**, 214-226.
90. Oliver, C., Mallet, V., Gendron, R.S., Reinharz, V., Hamilton, W.L., Moitessier, N. and Waldispuhl, J. (2020) Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic Acids Res*, **48**, 7690-7699.
91. Yan, Z., Hamilton, W.L. and Blanchette, M. (2020) Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. *Bioinformatics*, **36**, i276-i284.

Figure 1. Workflow of the RNANetMotif method. **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. **Step 3.** Calculate HVDM (Heterogeneous Value Difference Metric) distance matrix and construct similarity network of EKSeS. We display the folding and unfolding between sequence and secondary structure to describe class choices to deal with categorical attributes in HVDM distance. After unfolding, EKSeS (marked by dashed circles in the secondary structure) backed to primary sequence as gapped sequences with two k-mers falling into the four intervals $\{Ik, k=1,2,3,4\}$ corresponding to the coordinates. **Step 4.** Detect significant network modules and then identify intrinsic EKS motifs. **Step 5.** RNA 3D structure modeling via discrete molecular dynamics folding simulations and protein-RNA docking with simulation for validation.

Figure 2. Definition and classification of EKS. As displayed, there are three categories of EKS: opposite direction extensions, mixed-mode extensions and same direction extensions. Black edge is the backbone edges, the red edge corresponds to the base pair interaction between i and j , and green dashed edge is the possible base pair interactions (there may be one or two of them appear, in most cases, only one of them exists in every graph for pseudoknot free RNA structure).

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

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

Figure 6. Boxplot of atom counts calculated from modeled 3D structure of motifs and other regions for 4 RBPs.

Figure 7. Complex simulation ensemble of docking of LIN28B CSD domain and three RNA network motifs. MD ensembles (top) and snapshots (bottom) showing the stability of Protein-RNA interface and key amino acids forming phosphate (blue) and base (orange) contact. LIN28B CSD protein is shown in red and 100 nt RNA shown in cyan and blue, identified motifs are shown in yellow with specific sequence information.

RBP_Cell line	RNASeqMotif	BEAM	Graphprot	sshmm_shapes	sshmm_structures	common motifs
AKAP1_HepG2						CUG in hairpin loop CUG in internal loop
FAM120A_K562						
G3BP1_HepG2						AGAA in hairpin loop
IGF2BP1_K562						CAU in hairpin loop CUG in internal loop
LARP4_HepG2						CUG in hairpin loop
LIN28B_K562						AAGAA in hairpin loop
PABPC4_K562						
PUM2_K562						UGU in internal loop
RBM15_K562						CAG in internal loop
RPS3_HepG2						CAG in hairpin loop
SND1_HepG2						
UCL5_K562						AGAA in multiloop
UPF1_HepG2						CUG in internal loop
UPF1_K562						UCC in hairpin loop
YBX3_K562						
ZNF622_K562						AGAAG in multiloop

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