Methods 118–119 (2017) 73–81

Contents lists available at ScienceDirect

Methods

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

A combined sequence and structure based method for discovering enriched motifs in RNA from *in vivo* binding data

Maya Polishchuk [a,b], Inbal Paz [a], Refael Kohen [a], Rona Mesika [a], Zohar Yakhini [c,d], Yael Mandel-Gutfreund [a,c,*]

[a] Faculty of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel
[b] Vavilov Institute of General Genetics, Russian Academy of Science, Moscow 11933, Russia
[c] Faculty of Computer Science, Technion-Israel Institute of Technology, Haifa 32000, Israel
[d] School of Computer Science, Herzliya Interdisciplinary Center, Herzliya 46150, Israel

ARTICLE INFO

Article history:
Received 5 January 2017
Received in revised form 28 February 2017
Accepted 3 March 2017
Available online 6 March 2017

Keywords:
RNA binding proteins
RNA sequence and structure motifs
RNA secondary structure
Motif enrichment
Computational ranked based approach
CLIP-seq
SMARTIV

ABSTRACT

RNA binding proteins (RBPs) play an important role in regulating many processes in the cell. RBPs often recognize their RNA targets in a specific manner. In addition to the RNA primary sequence, the structure of the RNA has been shown to play a central role in RNA recognition by RBPs. In recent years, many experimental approaches, both *in vitro* and *in vivo*, were developed and employed to identify and characterize RBP targets and extract their binding specificities. *In vivo* binding techniques, such as CrossLinking and ImmunoPrecipitation (CLIP)-based methods, enable the characterization of protein binding sites on RNA targets. However, these methods do not provide information regarding the structural preferences of the protein. While methods to obtain the structure of RNA are available, inferring both the sequence and the structure preferences of RBPs remains a challenge. Here we present SMARTIV, a novel computational tool for discovering combined sequence and structure binding motifs from *in vivo* RNA binding data relying on the sequences of the target sites, the ranking of their binding scores and their predicted secondary structure. The combined motifs are provided in a unified representation that is informative and easy for visual perception. We tested the method on CLIP-seq data from different platforms for a variety of RBPs. Overall, we show that our results are highly consistent with known binding motifs of RBPs, offering additional information on their structural preferences.

1. Introduction

RNA binding proteins (RBPs) are essential for many processes in the cell, both in the nucleus and in the cytoplasm. Many RBPs recognize specific binding sites on their RNA target. These binding sites are usually characterized by specific short sequences, known as binding motifs. In addition to the primary RNA sequence, the structure of the RNA target is known to play a central role in guiding RBP-target recognition. It is well established that most RBPs prefer to bind their targets at single stranded regions [1]. However, some proteins, such as those possessing the double stranded RNA binding domain (dsRBD), e.g. Staufen, are known to bind specifically to dsRNA [2]. While it has been commonly believed that the dsRBDs recognize their RNA targets in a non-sequence specific manner, recent studies have shown that they recognize both sequence and structural determinants of the RNA [3]. In addition, RBPs belonging to other domain families have been shown to bind in a sequence specific manner to preferred RNA secondary structures, such as the yeast protein Vts1, which was experimentally verified to bind to a sequence motif within a loop of a hairpin structure [4].

In recent years, many high-throughput binding techniques have been developed to identify the binding preferences of RBPs. These technologies can be roughly divided into methods that measure protein-RNA binding *in vitro*, based on High Throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX), such as RNAcompete [5,6] and *in vivo* binding experiments, based on CrossLinking and ImmunoPrecipitation (CLIP). CLIP (HITS-CLIP) was originally introduced to identify the binding target of the neuronal specific RBP Nova in the mouse transcriptome [7]. Since then many different variants of the method have been developed and applied to a large number of RBPs in different cell types, attempting to increase the sensitivity and specificity of the methods. Among these methods are PAR-CLIP [8], iCLIP [9], eCLIP [10] and

* Corresponding author at: Faculty of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel.
E-mail address: yaelmg@tx.technion.ac.il (Y. Mandel-Gutfreund).

http://dx.doi.org/10.1016/j.ymeth.2017.03.003
1046-2023/

irCLIP [10]. Generally, the CLIP-based protocols start with UV irradiation of the cells to induce covalent crosslinks between RBPs and the RNAs, followed by immunoprecipitation of the bound protein-RNA complexes. Further the target sequences are extracted and sequenced using high-throughput sequencing. Finally, the sequences are mapped to the reference genome or transcriptome and analysed by dedicated bioinformatic analysis tools that are tailored to process the data resulting from the different CLIP methodologies. These dedicated tools are used to extract the binding sites, namely peak calling, and quantify their binding signals. For an extensive review on the different computational method for CLIP data analyses, see Uhl et al. in this special issue of Methods [11].

The next important step in analysing both *in vitro* and *in vivo* binding data is extracting the binding preferences of the proteins of interest, known as motif discovery. Over the years, motif discovery has attracted extensive research, resulting in hundreds of different tools, such as methods for discovering Transcription Factor binding motifs [12]. In addition, many methods have been developed to infer the binding preferences of RBPs from high-throughput RNA binding data (for an extensive review see [1]). Given the accumulating information from structural studies of protein-RNA, showing that the RNA structure plays a major role in protein-RNA recognition, many of the dedicated RNA motif-search algorithms consider both the primary sequence as well as the structural attributes of the RNA [1], mostly derived from RNA secondary structure predictions. MEMERIS was originally developed for extracting the binding preferences of splicing factors [13]. MEMERIS integrates secondary structure information (specifically single strandedness prediction) to the Expectation Minimization motif discovery algorithm, implemented in the popular motif discovery tool MEME [14]. The assumption behind MEMERIS is that the RBPs preferentially bind to ssRNA regions and thus it is designed to identify enriched motifs only in predicted single stranded regions. More recently, the same group has introduced GraphProt [15] for modelling the sequence and structure preferences of RBPs from either *in vivo* or *in vitro* data with no prior assumption regarding the binding preferences of the proteins. Different from MEMERIS, GraphProt is based on machine learning discriminative models trained on the information from bound versus the unbound data. Other approaches for modelling sequence and structure preferences of RBPs have been specifically designed to extract the binding preferences from *in vitro* data. RNAcontext [16] and the more recent method RCK [17] have been employed for extracting sequence and structure preferences from RNAcompete data [5,6]. In the latter methods, the RNA structure is predicted by RNA folding algorithms (such as RNAplfold [18]) and is represented as probabilities in the model, considering different types of paired and unpaired RNA conformations. As in GraphProt, RNAcontext and RCK do not require prior assumptions regarding the secondary structure preferences of the proteins and thus they can detect preferences of RBPs to bind in different structural contexts. Dao et al. developed aptaTRACE [19] for identifying sequence–structure motifs from HT-SELEX data, taking into consideration information from all rounds of SELEX selection, thus extending the current models discriminating bound from unbound data. The TEISER (Tool for Eliciting Informative Structural Elements in RNA) computation framework uses whole genome-wide measurements to extract enriched sequence–structure motifs of RBPs [20]. TEISER was successfully employed for discovering the structural preferences of the TARBP2 RBPs from CLIP data [21]. Recently, deep learning approaches that can incorporate information from different sources of data obtained by both *in vitro* and *in vivo* technologies have been introduced for predicting the binding specificities of RBPs. Currently, these methods have been applied for detecting sequence [22] and structure [23] binding preferences of RBPs, independently.

As more and more data is accumulating in the public databases from CLIP-seq binding experiments (e.g. DoRiNA [24], CLIPdb [25]), there is a strong need for bioinformatic tools that can be used for discovering the sequence and structure binding preferences from *in vivo* data. Here we present a dedicated method, named SMARTIV (Sequence and Structure Motif enrichment Analysis for Ranked RNA daTa generated from In-Vivo binding experiments) for extracting enriched motifs from *in vivo* high-throughput RNA binding data, combining sequence and secondary structure information. SMARTIV uses the numerical binding scores obtained from CLIP results and the predicted secondary structure of the sequences to generate a ranked list of sequences in a combined sequence and structure alphabet. Further, SMARTIV employs the DRIMUST algorithm to efficiently extract *k*-mers from ranked sequence data using suffix trees [26]. Finally, we provide a new motif representation that is informative and easy for visual perception. The extracted motifs contain both sequence and structural information concisely represented in a graphical logo using the eight symbol alphabet *A, C, G, U, a, c, g, u* (upper case for unpaired and lower case for paired nucleotides). We show that the combined sequence and structure motifs generated from the CLIP data are highly consistent with previously known sequence and structure binding preferences of the proteins.

2. Methods

2.1. Overview of the method

We present a *k*-mer based approach for efficient extraction of combined sequence and structure motifs of RBPs from a ranked list of RNA sequences, experimentally derived from *in vivo* high-throughput RNA binding assays. The method is currently designed for analysing processed CLIP-seq data, generated from CLIP databases, though it can be applicable for any kind of *in vivo* RNA binding data that can be ranked based on a given numeric score. As a first step, for a given RBP we extract the analysed bound sequences from the CLIP database and sort them by their reported binding score. Next, we map the sequences to their genomic location and extend the sequences to a defined length that is then submitted to the folding algorithm for predicting the secondary structure of each nucleotide in the RNA sequence. In the current implementation we employ the RNAsubopt algorithm for RNA folding [18]. Based on the secondary structure predictions, we extend the traditional 4-letter alphabet of the RNA to a new 8-letter alphabet, where each letter in the alphabet represents a specific nucleotide in a defined secondary structure. Notably, we consider only two possible states for defining a secondary structure preference: unpaired (single stranded) and paired (double stranded). We then apply the DRIMUST algorithm, implemented in our DRIMUST web server [26], for extracting the most significant short words (*k*-mers) that are highly enriched at the top of the list of the bound sequences, using the minimal Hypergeometric Statistics (mHG) [27]. Further, for each length *k* we cluster the significant *k*-mers selected from the 4-letter list and from the translated 8-letter list (separately for each list) and score the clusters based on the occurrences of the *k*-mers that constitute the cluster at the top of the list, where "top" is defined by the mHG statistics. Finally, we generate Position Weight Matrices (PWMs), representing the enriched motifs, by aligning the *k*-mers in each cluster. In addition, we estimate the *p-values* of the motifs generated from each cluster by correlating the predicted score of each sequence, calculated based on the PWM, to the binding score obtained from the original experiment. SMARTIV reports the PWMs with the best *p-values*, for the 4-letter and 8-letter alphabets independently. A flowchart describing the algorithm is shown in Fig. 1.

2.2. Methodology details

2.2.1. Data pre-processing

2.2.1.1. Sorting the CLIP results. As a first step, we extract the list of processed sequences from a given CLIP experiment or a CLIP database. When the list is provided as coordinates, we map the coordinates to the corresponding genome and extract the RNA sequences in FASTA format (based on the information provided regarding the genome assembly). Given that SMARTIV relies on the sequence ranking we consider only datasets for which a binding score is provided per each sequence in the processed list. We then sort the list by the binding score value in descending order (stronger binding scores are at the top of the list) after removing sequences shorter than a minimal length *lmin (*by default *lmin* = 13*)*. For efficiency, we select from the ranked list a total of 10,000 sequences in the following way: 1000 sequences from top of the list, and 9000 sequences from bottom of the list.

2.2.1.2. RNA secondary structure prediction. In an attempt to generate a combined sequence and structure alphabet, we need to match each sequence in the ranked list (derived in step 2.2.1.1) with its corresponding RNA structure. Given the dynamic nature of RNA, the strong influence of RBPs on RNA folding and the enormous challenges in obtaining the three dimensional structure of the entire cellular transcriptome, it is commonly accepted to consider RNA structure at the level of secondary structure. Thus, we seek to assign a predicted secondary structure to each nucleotide in the sequence list. Over the years many algorithmic approaches were developed for predicting RNA secondary structure (for review see [28]). More recently, high-throughput experimental approaches for inferring RNA secondary structure have also been developed [29]. However, experimental data on RNA structure is still limited to partial regions of the transcriptome of specific cell lines. Given the above, we chose to predict the RNA secondary structure using the RNAsubopt algorithm from the RNA Vienna package [18]. RNAsubopt calculates the RNA suboptimal structures with the lowest free energy, finally defining each nucleotide in the sequence as either paired or unpaired. The length of the input sequence for RNAsubopt is defined by a parameter *l*, ranging between 100 and 300 nucleotides, consistent with the accepted knowledge in the field for the optimal sequence length for folding RNA sequences. For each CLIP-sequence we retrieve flanking regions from the corresponding genome so that the total length of each sequence provided to RNAsubopt is equal to *l*. In cases where the original sequences are longer than *l*, we extract a sequence of length *l* from the center of the sequence. Subsequently, each nucleotide in the sequences data is assigned a predicted secondary structure (paired/unpaired).

2.2.1.3. Translating the sequences to a combined sequence and structure alphabet. The assigned secondary structure for each nucleotide in the original C:IP=sequence is retrieved and matched to the original sequence, finally keeping the original sequence length (excluding the flanking regions that were generated for folding purposes, as described in 2.2.1.2). In cases where the original sequence length is longer than *l* we keep the truncated sequences. Further we translate the sequences to an 8-letter alphabet *{a c g u A C G U}* where *a, c, g, u* correspond to paired nts *A, C, G, U*. Finally, we end up with two parallel sorted lists, the original ranked CLIP-seq data in the 4-letter alphabet *{A C G U}* and the folded CLIP-seq sequences in an 8-letter alphabet *{a c g u A C G U}*.

2.2.2. Extracting enriched k-mers from the ranked CLIP data

Similar to other *k*-mer based approaches for *de novo* motif identification (e.g. [17]), our algorithm is based on the assumption that binding motifs are derived from overrepresented sub-sequences of length *k* (*k*-mers) that occur more frequently in the bound sequences (as defined by the experimental assay). Here we extract *k*-mers from each ranked list, namely from the 4-letter alphabet list (original CLIP-seq data) and the translated sequence list in the 8-letter alphabet. To extract enriched *k*-mers we employ our *de novo* motif search algorithm [27], implemented in the DRIMUST webserver [26]. DRIMUST is a rank based approach for detecting imbalanced enriched motifs and thus is highly suitable for extracting enriched *k*-mers from CLIP data, in which the sequences are ranked according to a given binding score. The great advantage of DRIMUST over other algorithms for extracting overrepresented *k*-mers is that it searches the *k*-mers at the top of the input sequences list, where the top of the list is dynamically determined by the mHG statistics without a requirement to define bound versus unbound. For each *k*-mer DRIMUST assigns a statistical significance value using an mHG score, corrected for multiple testing, which is a tight bound to the *p-value* (*p-value* $\leq$ corrected mHG score) [27]. DRIMUST uses suffix trees for efficient enumeration of the candidate *k*-mers and produces the significant *k*-mers of different lengths *k* in a given range. While in DRIMUST the range of *k* in not limited, in SMARTIV we define the preferred range of *k*-mers that is most applicable for extracting motifs of RBPs. This range is provided as a parameter and can be changed by the users.

2.2.3. Generating adjusted PWMs from the detected enriched k-mers

In the next steps, the enriched *k*-mers from a selected range of length *k* are used to generate PWMs. The details of the PWM extraction procedure are given below (Sections 2.2.3.1, 2.2.3.2 and 2.2.3.3). Briefly, for each length *k* we cluster all significant *k*-mers generated by DRIMUST. We then sort the resulting clusters (generated for each length *k*) based on the occurrences of the enriched *k*-mers and represent each cluster as a matrix that is assigned a *p-value*, based on its correspondence to the original binding scores. As depicted in Fig. 1, the procedure is performed for all of the PWMs extracted from a given CLIP dataset, obtaining a list of sequence motifs (in 4-letter alphabet) and a list of combined sequence and structure motifs (in 8-letter alphabet) ranked by their *p-values*.

2.2.3.1. Clustering the k-mers. Clustering of the *k*-mers is performed for each length *k* separately using VSEARCH, a greedy centroid-based algorithm with an adjustable sequence similarity function [30,31]. Prior to the clustering, we sort the enriched *k*-mers based on the *p-value* obtained for each *k*-mer by the DRIMUST algorithm [26]. Briefly, the clustering process starts by selecting an initial *k*-mer with the lowest *p-value* from the *k*-mers list, which is then used as the cluster centroid. Subsequently, *k*-mers are added to the cluster if their similarity to the centroid is equal to or above a given threshold, where similarity is defined as number of matches between the *k*-mers divided by the alignment length. The process continues by selecting the next unclustered *k*-mer in the ranked list as the centroid for a new cluster and is repeated until all the *k*-mers are assigned to a cluster. Further, we align the *k*-mers in each cluster by conducting a semi- multiple sequence alignment of the centroid with all of the *k*-mers in the cluster, prohibiting internal gaps.

2.2.3.2. Building Position Weight Matrices. To generate a PWM from a given cluster we multiply each *k*-mer in the aligned cluster by the number of times the *k*-mer was found at the top of the list, as defined by the DRIMUST algorithm [26]. Further, we generate a PWM of dimension $b * a$, where *a* is the size of the alphabet (in our case 4 for sequence only alphabet and 8 for the combined sequence and structure alphabet) and *b* is the alignment length. For the graphical representation, we use a modified version of

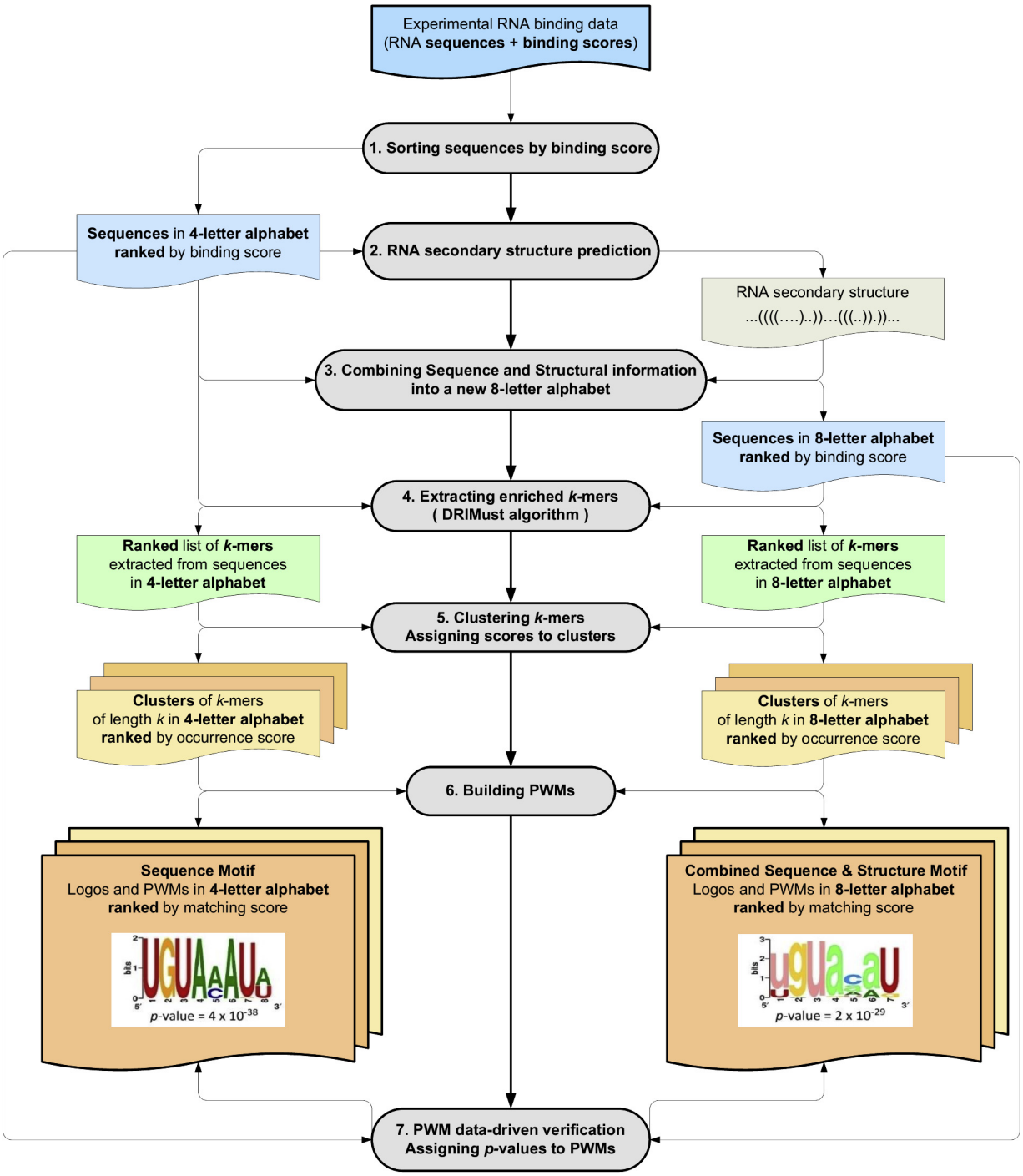


Fig. 1. A flowchart describing SMARTIV. The SMARTIV method takes as an input RNA binding data (e.g. CLIP-seq data). The output of the method is motif logos and corresponding PWM's ranked by the *p-value* of the combined sequence and structure motifs (in 8-letter alphabet) and sequence motifs (in a 4-letter alphabet). Each step of the algorithm is shown as a grey oval. The different steps of the algorithm are connected with bold arrows. Colored rectangles represent data sources: sequence data is in light blue, *k*-mers data is in light green, clusters of *k*-mers and resulting motifs are in shades peach colors. The input and output of each step of the algorithm are connected to the process via thin arrows. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the WebLogo algorithm [32], adjusted to present the PWMs for both the 4-letter alphabet and the new 8-letter alphabet (Fig. 1).

2.2.3.3. Assigning occurrence scores and p-values to PWMs. To select the best motifs for a given RBP, for each cluster we assign an *occurrence score,* which is defined as the total number of occurrences of *k*-mers (within the cluster) at the top of the list of the sorted CLIP sequences (in either original 4-letter alphabet or translated 8-letter alphabet). While the *occurrence score* can be used for ranking the clusters derived from a set of *k*-mers of a given length *k*, it is not applicable for comparing between clusters of *k*-mers of different lengths. Moreover, it cannot be used to compare PWMs generated from different CLIP datasets (either from the same or from different RBPs). To surmount these limitations, we assign a *p-value* to each PWM based on the match of the PWM to the original binding scores, derived from the CLIP data. To this end we scan each sequence in the ranked list against the PWMs that were derived from the *k*-mer clusters and we score each sequence based on its

match to the given PWM. The score for each individual sequence is defined as either the *max-score*, i.e. the *max score* for a sub-sequence of length *b* (number of columns in the PWM) in the sequence or the *sum-score*, i.e. the sum of scores for all overlapping sub-sequences, where the score for a sub-sequence is calculated as the sum of the log-odd scores over all positions in the sub-sequence. The background probabilities used to calculate the log odds score are defined as 0.25 and 0.125 for the 4-letter and 8-letter alphabet, respectively. Finally, the statistical significance of the correspondence between the original sequence binding scores (derived from the CLIP experiment) and the *match score* between the PWM and the sequence is estimated using the mmHG statistics, which evaluates the association between two ranked lists [33–35], assigning an FDR corrected *p-values* to each PWM.

2.3. Method implementation

The SMARTIV method has been implemented to a program running on unix operating system via simple terminal-based command line access and a user-friendly webserver. The input for the SMARTIV method is a CLIP-seq processed file including the list of target sequences and their corresponding binding scores sorted in descending order based on the score. The target sequences can be provided as either coordinates (in BED format) or as sequences in FASTA format. Notably, if structural information is available, either from other secondary structure prediction algorithms or experimental data, users can provide the ranked list of sequences in 8-letter alphabet. When running SMARTIV, the users can select the range of *k*-mers for generating the PWMs or choose the default range (4–6). The output of SMARTIV is a list of predicted combined sequence and structure motifs in the 8-letter alphabet as well as in the traditional 4-letter alphabet, ranked by their *p-value*. Both the combined motifs and the sequence motifs are provided in a graphical format (WebLogo format) and as a PWM with their corresponding *p-values*. In addition, SMARTIV provides a detailed list of the enriched *k*-mers that were used to generate the PWM (calculated by DRIMUST [26]). The SMARTIV source code is available upon request. The webserver is accessible through the website http://smartiv.technion.ac.il.

3. Results and discussion

3.1. Datasets

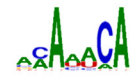
We have tested our method on a variety of CLIP datasets for many different RBPs generated by different CLIP methods including HITS-CLIP [36], PAR-CLIP [8] and iCLIP [9]. The datasets were extracted from two dedicated databases of CLIP data: DoRiNA [24] and CLIPdb [25]. An important feature of our motif prediction algorithm is that it calculates the *k*-mer enrichment from a list of sequences, sorted by their binding scores, without a requirement to arbitrarily split the datasets to bound and unbound sets. Furthermore, SMARTIV relies on the ranking of the CLIP binding scores for evaluating the significance of the PWMs generated from the data. Thus, to test our method we have chosen from the databases only datasets for which binding scores are available. Here we present results for a selected set of representative RBPs from different RNA binding families, for which their sequence binding motifs (extracted from either *in vitro* or *in vivo* studies) have been previously reported. Among these proteins we included RBPs for which their structural preferences have been predicted by either motif detection algorithms or inferred from structural or biochemical experiments. Detailed information on the data sources of the CLIP data of the selected proteins is given in Table S1.

3.2. Combined sequence and structure motifs are consistent with sequence motifs

In Fig. 2 we present the motifs generated by SMARTIV for eight selected RBPs, illustrating the motifs in 8-letter alphabet (generated from the translated lists) and in 4-letter alphabet (generated from the sequence lists). Conventional upper case letters in the graphical logos represent nucleotides predicted to be in unpaired (single stranded) RNA conformations while letters in lower case represent nucleotides predicted to be paired. For better perception we use different color shading for upper and lower case letters (see color bar under the Table). As demonstrated in Fig. 2, in the majority of cases the combined sequence and structure motifs (specifically for ELAVL1, IGF2BP2, QKI, SRSF1, TIA1) are very similar to the sequence motifs and are all in upper case and dark shading letters, indicating the preference of the RBPs to be found in ssRNA. These results are consistent with the well-established knowledge in the field that most RBPs tend to bind to single stranded binding sites on RNA. Notably, the strong agreement between the sequence and structure motifs and the sequence only motifs shown in Fig. 2 is completely not trivial as the enriched *k*-mers that were used to build the 8-letter motifs are selected from the independent list of translated sequences. Interestingly, while as expected the *occurrence score* for the clusters that is generated from *k*-mers in the 8-letter alphabet is usually lower than for the clusters generated from *k*-mers in 4-letter alphabet, the *p-values* of the PWMs, representing the match to the experimental binding score, are highly consistent (Fig. 2). These results reveal that the *k*-mers from which the motifs were derived are preferentially selected from stretches of sequences that are predicted by the folding algorithm as unpaired, reinforcing that the RBPs preferentially recognize their binding sites in a ssRNA context. Overall, the motifs generated by SMARTIV (both the 8-letter and the 4-letter motifs) are usually consistent with the motifs reported for these RBPs in the literature [6,8,16,37–41]. For comparison, in Fig. 2 we present the motifs generated from the same experiment by other algorithms or when unavailable from another CLIP-seq experiment that was conducted for the same RBP. In addition, the motif from the RNA compendium, predicted by RNAcontext [6] from in vitro data, is provided. Notably, in Fig. 2 we present only the most significant motif among all the motifs generated from *k*-mers of different lengths *k*, i.e. the motif that is most consistent with the experimental binding data. Nevertheless, as shown in Table S2, the motifs predicted from *k*-mers of different lengths (within a given range of *k*) are usually highly consistent between each other and are generally in accordance with previously reported motifs.

3.3. Combined sequence and structure motifs demonstrate known and novel structural preferences for RNA binding proteins

As aforementioned, during the last decade, several methods have been developed for detecting sequence and structure motifs of RBPs. However, the majority of these methods were developed and tested on *in vitro* data. To assess the 8-letter motifs predicted by SMARTIV from the CLIP-seq we compared our predicted motifs to motifs generated for the respective RBPs by the GraphProt state-of-the-art algorithm [15]. In Table S3 we present motifs for seven RBPs for which GraphProt motifs (for paired/unpaired preferences) were available in the literature [15]. As shown in Table S3, the 8-letter motifs predicted by SMARTIV are generally in agreement with the motifs predicted by GraphProt. Note that the great advantage of our 8-letter motifs over the GraphProt motifs is that SMARTIV represents the sequence and structural information in a unified representation, thus presenting the likelihood of each nucleotide in the motif to be in either paired or unpaired conformation. On the contrary, GraphProt presents the sequence and structure

Protein Info		SMARTIV Motif Prediction		Known motifs	
Name	**Dataset**	**Combined Sequence & Structure Logo**	**Sequence Logo**	***in vivo***	***in vitro***
ELAVL1	*human* PAR-CLIP				
		7.41×10^{-31}	3.29×10^{-32}	[16, 39]	[6]
FMR1	*human* PAR-CLIP				
		1.03×10^{-232}	0	[37]	[6]
IGF2BP2	*human* PAR-CLIP				
		6.41×10^{-89}	0	[8]	[6]
PUM2	*human* PAR-CLIP				 *drosophila*
		2.31×10^{-29}	4.24×10^{-38}	[8]	[6]
QKI	*human* PAR-CLIP				
		2.07×10^{-15}	5.58×10^{-19}	[8]	[6]
SRSF1	*mouse* HITS-CLIP			 *human*	 *human*
		1.15×10^{-70}	4.7×10^{-60}	[16, 41]	[6]
TDP43	*human* iCLIP				
		0	0	[38]	[6]
TIA1	*human* iCLIP				
		1.67×10^{-34}	4.56×10^{-58}	[40]	[6]

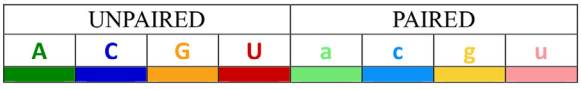


Fig. 2. Combined sequence and structure motifs for eight selected proteins. SMARTIV results for eight selected proteins, including sequence and structure motifs (in 8-letter alphabet) and sequence motifs (in 4-letter alphabets), are presented. The *p-value* representing the match of the PWM to the experimental data is shown under each motif. The published motifs generated from *in vivo* or *in vitro* data are presented. A letter and color code for the logo representations are given under the table.

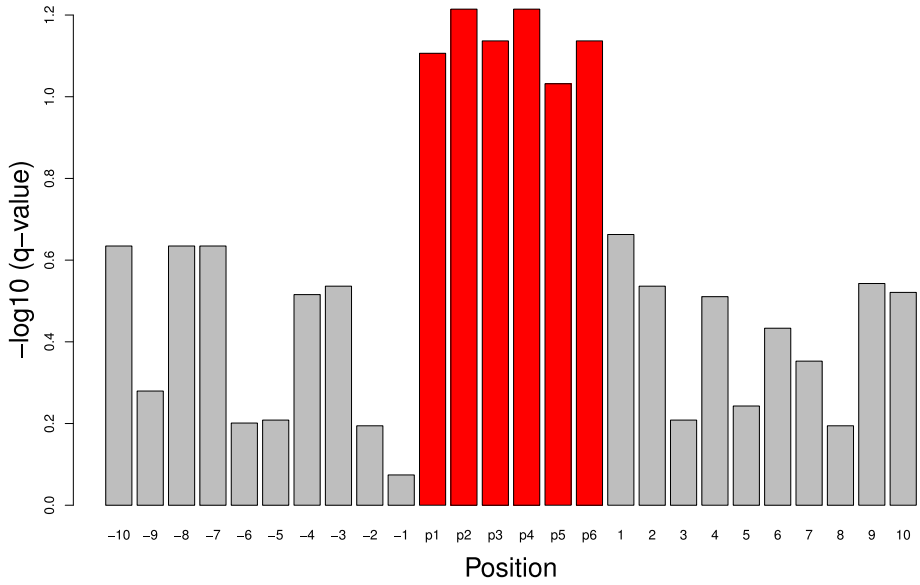


Fig. 3. Difference in the conservation of the sequences possessing the PUM2 sequence and structure enriched motif versus the sequence only motif. Bars represent the FDR corrected *–log (p-value)* for the Mann Whitney Wilcoxon test, comparing the conservation of the sequences possessing the enriched *k*-mers of the selected PWM from the 8-letter alphabet to the enriched *k*-mers of the selected PWM from the 4-letter alphabet. Sequence conservation was calculated for each position of the *k*-mer and ten nucleotides upstream and downstream. Conservation values for the original CLIP sequences were retrieved from the UCSC phyloP table for placental mammalians.

preferences in separate logos from which the correspondence between the sequence and the structure preferences of individual nucleotides can only be deduced from visual inspection of the two independent logos. As for example, in the case of the PUM2 motif, while GraphProt predicts that all the positions in the motif have similar probabilities to be in either paired or unpaired regions, the motif predicted for PUM2 by SMARTIV shows a strong preference for the 5′ part of the motif, specifically the core GUA, to be in a paired RNA conformation. The binding preference of the Pumilio RBP family has been puzzling the protein-RNA field for many years. While the structural data clearly show that the members of the family bind to well-defined ssRNA sequences [42], biochemical and genomics studies suggest that the proteins have an important functional role in miRNA binding by inducing structural changes in non-accessible dsRNA regions [43,44]. To evaluate SMARTIV 8-letter predicted motif for the human PUM2 protein, we compared the sequence conservation of the *k*-mers belonging to the best cluster extracted from the 8-letter alphabet to the conservation of the *k*-mers that consist the best cluster from the 4 letter alphabet. To this end we extracted the sequences that consist the *k*-mers of length 6 that were used to derive the best PUM2 motif (PWM). Note the overall length of the motif generated from the *k*-mers is not necessarily of length *k*. In the latter case the PUM2 motif generated from *k*-mers of length 6 is of length 7. We further calculated for each position in the mapped 6-mer and ten flaking positions from each side, their conservation in placental mammalians, retrieved from the UCSC phyloP conservation table [45]. Consequently we compared the conservation values for the sequences consisting the 6-mers and flanking regions from the 4-letter and 8-letter alphabet and compared between the two groups using the Man-Whitney Wilcoxson *U* test, applied for each position in the motif and the flaking regions independently. Overall, the conservation values of the sequences possessing the enriched *k*-mers from the 8-letter alphabet were higher than those possessing the enriched *k*-mers from the 4-letter sequence only alphabet. Strikingly, as shown in Fig. 3, significant differences were found merely for the positions within the motif. The conservation results strongly support that the subset of PUM2 motifs, found within regions predicted to be in paired conformation, tend to be more conserved than the set of sequences containing the sequence motif only. These results support SMARTIV prediction that the well-defined PUM2 core sequence motif UGUA tends to reside in partial dsRNA regions. As shown in Fig. 2, another RBP predicted by SMARTIV to be partially in a paired structure is TDP-43. TDP-43 is an RBP, possessing the RNA Recognition Motif (RRM), known to be involved in several neurodegenerative diseases, including amyotrophic lateral sclerosis (ALS) [38]. Consistent with GraphProt and many other motif detection methods, SMARTIV predicts that the binding motif of TDB43 consists of a stretch of UG repeats. Nevertheless, different than GraphProt, we predict that only the central UG di-nucleotide is in an unpaired conformation while the flanking repeats are predicted to be in paired conformation. Here again our prediction is not in agreement with the crystal structure of the TDP-43/RNA complex demonstrating that TDB-43 binds to a single stranded UG stretch [46]. Based on the highly consistent motifs we obtained from *k*-mers of different lengths (Table S1 and other results for different ranges of *k*, not shown), we suggest that TDP43, which is also known to bind DNA, may have a tendency to recognize the UG motif within weak RNA hairpin structures while finally binding the RNA in a single stranded conformation.

4. Conclusions

In this paper we present a new method named SMARTIV, for discovering combined sequence and structure motifs for RBPs from *in vivo* binding data generated from different CLIP-based methods. SMARTIV is available as a source code for download and as a web-server. The SMARTIV algorithm has several advantages over other recently developed methods for extracting sequence and structure motifs from high-throughput RNA binding data. Similar to other *k*-mer based motif detection algorithms, SMARTIV generates the motifs from sets of sub-sequences, found to be enriched in the data. However, while most other methods require splitting the data artificially to bound and unbound datasets, our method, considers the entire information from the dataset when ranked by the reported binding scores. Clearly SMARTIV results depend on the algorithm used to assign the binding score per sequence,

nevertheless we find that it is usually robust to different scoring systems. Moreover, the method is not restricted to a defined length of *k*-mers and can efficiently extract motifs from a large range of *k*-mers, finally choosing the motifs that are best correlated with the original experimental data. Notably, given that the motifs (PWMs) are generated from the enriched *k*-mers and are not directly extracted from the data, the algorithm does not estimate the statistical significance of the enriched motifs. Nevertheless, for each PWM we provide a *p-value* that represents its correlation with the experimental binding scores. The motif *p-values* are then used by the algorithm to select the best motifs. The great advantage of SMARTIV is that it generates combined motifs that represent the preference of each nucleotide to be in a paired or an unpaired RNA region in a simple and highly intuitive graphical manner. Importantly, while the combined sequence and structure motifs clearly depend on the folding algorithm used to predict the secondary structure of the RNA, known to be very noisy, only sub-sequences that are consistently found in the same RNA conformation (i.e. enriched *k*-mers in the translated sequences) are selected to generate the final motifs. Moreover, while currently our method relies on information from predicted secondary structure, it can be easily adapted to extract motifs from experimental folding data, once such data is available for the entire transcriptomes of the relevant cell lines or tissues for which the experiments were conducted. Finally, SMARTIV is extremely fast. On average, we process one CLIP dataset in approximately 3–4 min on an Intel Core i7-2600 CPU @ 3.40 Ghz * 4 and 32 Gb memory.

Funding

This work supported in part at the Technion by the Lady Davis fellowship granted to MP.

Acknowledgment

We would like to thank Limor Leibovich who developed the DRIMUST algorithm and provided access to the source code and to Leon Anavy for adjusting the DRIMUST code to SMARTIV. Thanks to Hagay Enav for contributing with the conservation analysis.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2017.03.003.

References

[1] X.C. Li, H. Kazan, H.D. Lipshitz, Q.D. Morris, Finding the target sites of RNA-binding proteins, Wiley Interdiscip. Rev. RNA 5 (1) (2014) 111–130.
[2] A. Ramos, S. Grunert, J. Adams, D.R. Micklem, M.R. Proctor, S. Freund, M. Bycroft, D. St Johnston, G. Varani, RNA recognition by a Staufen double-stranded RNA-binding domain, EMBO J. 19 (5) (2000) 997–1009.
[3] G. Masliah, P. Barraud, F.H. Allain, RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence, Cell. Mol. Life Sci. 70 (11) (2013) 1875–1895.
[4] T. Aviv, Z. Lin, G. Ben-Ari, C.A. Smibert, F. Sicheri, Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p, Nat. Struct. Mol. Biol. 13 (2) (2006) 168–176.
[5] D. Ray, H. Kazan, E.T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B.J. Blencowe, Q. Morris, T.R. Hughes, Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins, Nat. Biotechnol. 27 (7) (2009) 667–670.
[6] D. Ray, H. Kazan, K.B. Cook, M.T. Weirauch, H.S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L.H. Matzat, R.K. Dale, S.A. Smith, C. A. Yarosh, S.M. Kelly, B. Nabet, D. Mecenas, W. Li, R.S. Laishram, M. Qiao, H.D. Lipshitz, F. Piano, A.H. Corbett, R.P. Carstens, B.J. Frey, R.A. Anderson, K.W. Lynch, L.O. Penalva, E.P. Lei, A.G. Fraser, B.J. Blencowe, Q.D. Morris, T.R. Hughes, A compendium of RNA-binding motifs for decoding gene regulation, Nature 499 (7457) (2013) 172–177.
[7] D.D. Licatalosi, A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A.C. Schweitzer, J.E. Blume, X. Wang, J.C. Darnell, R.B. Darnell, HITS-CLIP yields genome-wide insights into brain alternative RNA processing, Nature 456 (7221) (2008) 464–469.
[8] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano Jr., A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell 141 (1) (2010) 129–141.
[9] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D.J. Turner, N.M. Luscombe, J. Ule, ICLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution, Nat. Struct. Mol. Biol. 17 (7) (2010) 909–915.
[10] E.U. Van Nostrand, G.U. Pratt, A.A. Shishkin, C.U. Gelboin-Burkhart, M.U. Fang, B.U. Sundararaman, S.U. Blue, T.U. Nguyen, C. Surka, K.U. Elkins, R.U. Stanton, F. Rigo, M. Guttman, G.U. Yeo, Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP), Nat. Methods 13 (6) (2016) 508–514.
[11] M.G. Uhl, T.G. Houwaart, G.I. Corrado, P.R.G. Wright, R. Backofen, Computational analysis of CLIP-seq data, Methods (2017).
[12] M.T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T.R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H.J. Bussemaker, Q.D. Morris, M.L. Bulyk, G. Stolovitzky, T.R. Hughes, Evaluation of methods for modeling transcription factor sequence specificity, Nat. Biotechnol. 31 (2) (2013) 126–134.
[13] M.G. Hiller, R. Pudimat, A. Busch, R. Backofen, Using RNA secondary structures to guide sequence motif finding towards single-stranded regions, Nucleic Acids Res. 34 (17) (2006) e117.
[14] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proc. Int. Conf. Intell. Syst. Mol. Biol. 2 (1994) 28–36.
[15] D. Maticzka, S.J. Lange, F. Costa, R. Backofen, GraphProt: modeling binding preferences of RNA-binding proteins, Genome Biol. 15 (1) (2014) R17.
[16] H. Kazan, D. Ray, E.T. Chan, T.R. Hughes, Q. Morris, RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins, PLoS Comput. Biol. 6 (2010) e1000832.
[17] Y. Orenstein, Y. Wang, B. Berger, RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data, Bioinformatics 32 (12) (2016) i351–i359.
[18] R. Lorenz, S.H. Bernhart, C. Honer, Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, ViennaRNA Package 2.0, Algorithms, Mol. Biol. 6 (2011) 26.
[19] P. Dao, J. Hoinka, M. Takahashi, J. Zhou, M. Ho, Y. Wang, F. Costa, J.J. Rossi, R. Backofen, J. Burnett, T.M. Przytycka, AptaTRACE elucidates RNA sequence-structure Motifs from selection trends in HT-SELEX experiments, Cell Syst. 3 (1) (2016) 62–70.
[20] H. Goodarzi, H.S. Najafabadi, P. Oikonomou, T.M. Greco, L. Fish, R. Salavati, I.M. Cristea, S. Tavazoie, Systematic discovery of structural elements governing stability of mammalian messenger RNAs, Nature 485 (7397) (2012) 264–268.
[21] H.U. Goodarzi, S.F. Tavazoie, S. Tavazoie, TARBP2 binding structured RNA elements drives metastasis, Cell Cycle 13 (18) (2014) 2799–2800.
[22] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (8) (2015) 831–838.
[23] M. Pietrosanto, E. Mattei, M.c.u.i. Helmer-Citterich, F. Ferre, A novel method for the identification of conserved structural patterns in RNA: from small scale to high-throughput applications, Nucleic Acids Res. 44 (18) (2016) 8600–8609.
[24] K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, A. Akalin, DoRiNA 2.0-upgrading the doRiNA database of RNA interactions in post-transcriptional regulation, Nucleic Acids Res. 43 (Database issue) (2015) D160–D167.
[25] Y.C. Yang, C. Di, B. Hu, M. Zhou, Y. Liu, N. Song, Y. Li, J. Umetsu, Z.J. Lu, CLIPdb: a CLIP-seq database for protein-RNA interactions, BMC Genomics 16 (2015) 51.
[26] L. Leibovich, I. Paz, Z. Yakhini, Y. Mandel-Gutfreund, DRIMust: a web server for discovering rank imbalanced motifs using suffix trees, Nucleic Acids Res. 41 (Web Server issue) (2013) W174–W179.
[27] E. Eden, D. Lipson, S. Yogev, Z. Yakhini, Discovering motifs in ranked lists of DNA sequences, PLoS Comput. Biol. 3 (3) (2007) e39.
[28] R.r.t.u.a.a. Lorenz, M.T.E.a.m.w.u.a.a. Wolfinger, A.a.t.u.a.a. Tanzer, I.L. Hofacker, Predicting RNA secondary structures from sequence and probing data, Methods 103 (2016) 86–98.
[29] I.M. Silverman, N.D. Berkowitz, S.J. Gosai, B.D.b.s.u.e. Gregory, Genome-Wide Approaches for RNA Structure Probing, Adv. Exp. Med. Biol. 907 (2016) 29–59.
[30] T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahe, VSEARCH: a versatile open source tool for metagenomics, PeerJ 4 (2016) e2584.
[31] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26 (19) (2010) 2460–2461.
[32] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (6) (2004) 1188–1190.
[33] I. Steinfeld, R. Navon, R. Ach, Z. Yakhini, MiRNA target enrichment analysis reveals directly active miRNAs in health and disease, Nucleic Acids Res. 41 (3) (2013) e45.
[34] D.I. Cohn-Alperovich, A.I. Rabner, I. Kifer, Y.I. Mandel-Gutfreund, Z. Yakhini, Mutual enrichment in aggregated ranked lists with applications to gene expression regulation, Bioinformatics 32 (17) (2016) i464–i472.
[35] L. Leibovich, Z. Yakhini, Mutual enrichment in ranked lists and the statistical assessment of position weight matrix motifs, Algorithms Mol. Biol. 9 (1) (2014) 11.

[36] R.B. Darnell, HITS-CLIP: panoramic views of protein-RNA regulation in living cells, Wiley Interdiscip. Rev. RNA 1 (2) (2010) 266–286.
[37] M. Ascano Jr., N. Mukherjee, P. Bandaru, J.B. Miller, J.D. Nusbaum, D.L. Corcoran, C. Langlois, M. Munschauer, S. Dewell, M. Hafner, Z. Williams, U. Ohler, T. Tuschl, FMRP targets distinct mRNA sequence elements to regulate protein expression, Nature 492 (7429) (2012) 382–386.
[38] C. Colombrita, E. Onesto, F. Megiorni, A. Pizzuti, F.E. Baralle, E. Buratti, V. Silani, A. Ratti, TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells, J. Biol. Chem. 287 (19) (2012) 15635–15647.
[39] F.B. Gao, C.C. Carson, T. Levine, J.D. Keene, Selection of a subset of mRNAs from combinatorial 3' untranslated region libraries using neuronal RNA-binding protein Hel-N1, Proc. Natl. Acad. Sci. U.S.A. 91 (23) (1994) 11207–11211.
[40] I.U. Lopez de Silanes, S. Galban, J.L. Martindale, X. Yang, K. Mazan-Mamczarz, F. E. Indig, G. Falco, M. Zhan, M. Gorospe, Identification and functional outcome of mRNAs associated with RNA-binding protein TIA-1, Mol. Cell. Biol. 25 (21) (2005) 9520–9531.
[41] R. Tacke, Y. Chen, J.L. Manley, Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer, Proc. Natl. Acad. Sci. U.S.A. 94 (4) (1997) 1148–1153.
[42] X. Wang, J. McLachlan, P.D. Zamore, T.M. Hall, Modular recognition of RNA by a human pumilio-homology domain, Cell 110 (4) (2002) 501–512.
[43] M. Kedde, M. van Kouwenhove, W. Zwart, J.A. Oude Vrielink, R. Elkon, R. Agami, A Pumilio-induced RNA structure switch in p27–3' UTR controls miR-221 and miR-222 accessibility, Nat. Cell Biol. 12 (10) (2010) 1014–1020.
[44] L. Leibovich, Y. Mandel-Gutfreund, Z. Yakhini, A structural-based statistical approach suggests a cooperative activity of PUM1 and miR-410 in human 3'-untranslated regions, in: Silence 1 (1) (2010) 17.
[45] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome Res. 15 (8) (2005) 1034–1050.
[46] P.H. Kuo, C.H. Chiang, Y.T. Wang, L.G. Doudeva, H.S. Yuan, The crystal structure of TDP-43 RRM1-DNA complex reveals the specific recognition for UG- and TG-rich nucleic acids, Nucleic Acids Res. 42 (7) (2014) 4712–4722.