Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1

Xin Wang[1,4], Liran Juan[1,5], Junjie Lv[4], Kejun Wang[4], Jeremy Sanford[6] and Yunlong Liu[1,2,3,§]

[1]Center for Computational Biology and Bioinformatics, [2]Department of Medical and Molecular Genetics, [3]Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN 46202, United States

[4]College of Automation, Harbin Engineering University, Harbin, Heilongjiang 150001, China

[5]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

[6]Department of Molecular, Cellular and Developmental Biology, University of California Santa Cruz, Santa Cruz, California 95064, United States

Abstract - *RNA-binding proteins (RBPs) play diverse roles in eukaryotic RNA processing. Despite their pervasive functions in coding and non-coding RNA biogenesis and regulation, elucidating the specificities that define protein-RNA interactions remains a major challenge. Here, we describe a novel model-based approach — RNAMotifModeler to identify binding consensus of RBPs by integrating sequence features and RNA secondary structures. Using RNA sequences derived from Cross-linking immunoprecipitation (CLIP) followed by high-throughput sequencing for SRSF1 proteins, we identified a purine-rich octamer 'AGAAGAAG' in a highly single-stranded RNA context, which is consistent with previous knowledge. The successful implementation on SRSF1 CLIP-seq data demonstrates great potential to improve our understanding on the binding specificity of RNA binding proteins.*

Keywords: protein-RNA binding, RNA secondary structure, motif, SRSF1, particle swarm optimization

1 Introduction

RNA-binding proteins (RBPs) are implicated in virtually every step of post-transcriptional gene expression including pre-mRNA splicing, RNA editing and polyadenylation [1]. These proteins possess a diverse array of structurally and functionally distinct RNA-binding domains such as RNA recognition motifs (RRM), KH domains, RGG boxes, zinc finger, double-stranded RNA-binding domain, etc [1]. Although the structures of many RNA binding domains have been solved at high resolution, establishing the sequence and RNA-structural determinants to binding specificity remains largely unexplored.

Several methods for elucidating the specificity of protein-RNA interactions enable rapid advances in our understanding of RBP functions. One recent innovation is the Cross-Linking ImmunoPrecipitation (CLIP). CLIP exploits photoreactive residues in RNA and polypeptides to generate covalently linked complexes. Because UV irradiation does not induce protein-protein cross-links CLIP is thought to be more specific than other IP based assays for protein-RNA interactions. CLIP was successfully applied to identify mRNA targets of the NOVA protein, a neural splicing factor associated with paraneoplastic opsoclonus myoclonus ataxia (POMA) [2-4]. Coupling CLIP with next-generation high-throughput sequencing technology, known as CLIP-seq or HITS-CLIP, provides a cost-efficient method to increase the sensitivity of the assay by surveying the RNA landscape on a more global scale. Several groups have successfully implemented CLIP-seq analysis of NOVA, SRSF1, fox2 and PTB proteins in mammalian systems [2, 5-7]. Both *MEME* and Z-score statistics have been used to reveal consensus binding motifs that are overrepresented in CLIP-Seq data [2, 6]. Although Z-score statistics may be able to find out the overrepresented sequence motifs, it does not consider the degenerated feature of the binding specificities of RBPs. MEME-based method is well known to be an excellent tool for cases only regarding sequence specificity [8]. Neither of these approaches can ascertain the roles of RNA secondary structure in establishing the context of the protein-RNA interaction. Hiller et al. extended MEME by adding a pre-computing procedure to measure single-strandedness of RNA sequence as *a priori* information to guide the motif search. They demonstrated that their model, *MEMERIS*, is able to identify binding motifs located in single-stranded regions with applications to both artificial and biological data [9]. Recently, Kazan et al. proposed *RNAcontext* for learning both sequence and structural binding preferences of RNA-binding proteins [10].

Here we describe a model-based approach—*RNAMotifModeler* to evaluate protein-RNA interactions using a retained binding affinity ratio, which is considered to be affected by two major factors—sequence degeneracy and RNA secondary structure deviation. *RNAMotifModeler* incorporates predicted unpaired probability of each nucleotide in the protein-RNA binding regions; such probability is derived from RNA secondary prediction algorithms (e.g. *RNAfold* [2]) based on the nucleotide compositions of the neighbouring flanking sequences. This strategy is different

from RNAContext, which uses predicted RNA secondary structures as input such as 'Paired', 'Hairpin Loop', 'Unstructured' or 'Miscellaneous'. Unlike MEMERIS, RNAMotifModeler uses the base-pairing probability for each nucleotide rather than the entire sequence (PU or EF values) [3]. For each binding instance, RNAMotifModeler defines a score that evaluates the consensus binding site within an optimal structural context, and aims at searching for an optimal RNA sequence-structural consensus for an RNA binding protein. These features enhance our ability to calculate and estimate the sequences that yield the highest binding affinity for a specific RBP.

We tested RNAMotifModeler on CLIP-seq data that profile the transcriptome-wide binding pattern of SRSF1, serine/arginine-rich splicing factor 1 [4]. The sequence features of the binding motifs is consistent with the experimentally defined *cis*-acting RNA elements recognized by SRSF1 [5]. Interestingly, the prediction suggests that the second and fifth bases of SRSF1 octamer motif have stronger sequence specificities, but lower p-values of unpaired probabilities, while the third, fourth, sixth and seventh bases are more significantly to be single-stranded, but have less sequence specificities. Therefore, we hypothesize that the sequence and structure specificities are both required and are playing complementary roles during binding site recognition of SRSF1.

2 Results

SRSF1 is an essential splicing factor with multiple roles in post-transcriptional gene expression [6]. SRSF1 is also a potent proto-oncogene and implicated in maintaining genome stability [7]. Moreover, loss of SRSF1 binding sites by mutations linked to genetic diseases can induce aberrant patterns of pre-mRNA splicing [4]. Thus considerable effort has been focused on defining the binding specificity and RNA targets of SRSF1. Here we report a novel model-based approach intended to examine the contributions of structural and sequence elements in RNA fragments co-purified with SRSF1 by CLIP.

2.1 Workflow of RNAMotifModeler

The first step of *RNAMotifModeler* is to do data preparations. In the present study, 904 positive gold standard sequences were selected from commonly targeted regions across three out of four samples in our previous SRSF1 CLIP-seq experiments [4]. The same number of negative sequences were randomly picked from non-SRSF1-targeted regions falling in the same genomic category (exonic, intronic, intergenic, etc) as their positive counterparts. Base pairing probabilities of each nucleotide to its neighbours were subsequently predicted by RNAfold [2] (ViennaRNA package, version 1.8.5) for both positive and negative gold standard sequences.

Our next step, as shown in Fig. 1, is to identify sequence-structural consensus using gold standard sequences and corresponding base pairing probabilities derived from RNAfold. We took an iterative approach that alternates between: 1) optimization of parameters specifying sequence degeneracy and structural context given a reference motif (the optimal binding sequence), and 2) searching for optimal reference motif given the estimated parameters by evaluation of each motif candidate's contribution to binding affinities of positive gold standard sequences (more details in Methods). The above two steps will be repeated until a convergence when the starting motif candidate makes the most contribution to binding affinities.

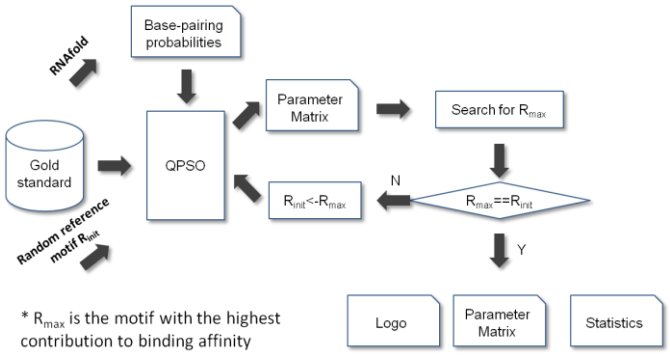

Fig. 1. Workflow of RNAMotifModeler

Finally, RNAMotifModeler outputs the converged reference motif, optimal parameters, statistical evaluation such as the AUC scores. The AUC scores are measured by the area under the ROC (Receiver Operating Characteristic) curves derived from predictions of gold standard sequences being bound by SRSF1 proteins using the predicted parameters. In order to predict binding sites of SRSF1 proteins, we pick the sequence binding affinity yielding the maximal prediction accuracy as a cutoff score. Based on the predicted reference motif and corresponding parameters, positive gold-standard sequences can be scanned to find all potential binding sites with binding affinities higher than the cutoff score. These binding sites can be further used to create a sequence consensus logo and transformed to positional weight matrix, which is much more widely used.

2.2 Convergence of SRSF1 consensus motif searching

We call the converging path from a starting motif candidate to the final consensus motif a *motif searching pathway*. This graph provides a visual demonstration on the pathways through which the reference motifs are determined. To have a global overview of the convergence, motif searching pathways for all motif candidates are organized together to form a *motif searching graph*. In the particular case of hexamer predictions for SRSF1, all 4096 motif candidates converge to a short list of candidates (Fig. 2). All motif candidates converge within three iterations, of which 85.7% converge after the first iteration. AGAAGA, AAGAAG and GAAGAA are top three hexamers with the highest in-degrees, responsible for 99.7% of all motif

candidates (Table 1). The other twelve reference motifs are closely related to these three motifs, only with one or two sequence alterations. It is also noted that nearly an equal number of motif candidates converge to each one of the top three reference motifs. More interestingly, these hexamers share a core of 'AAGA' indicating that they may be adjacent to each other in RNA fragments.

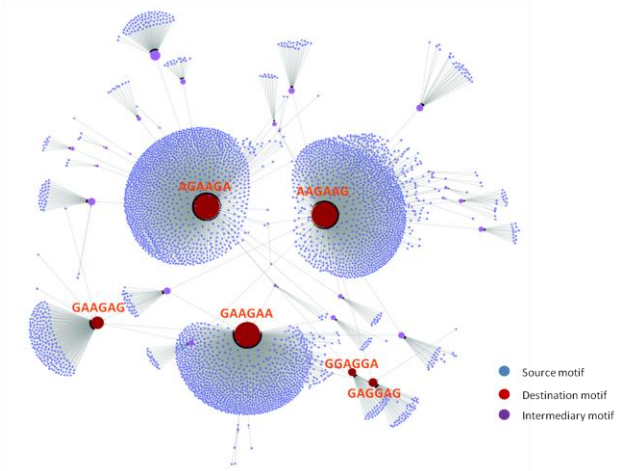

Fig. 2. Motif searching graph. Source, intermediary and destination motifs are denoted by nodes colored in blue, purple and red, respectively. The size of node is proportional to its in-degree. Arrows between nodes indicate converging directions. This figure demonstrates the fast convergence of the vast majority of motif candidates using the Quantum Particle Swarm Optimization algorithm.

Table 1. Converged motifs and corresponding numbers of source motifs

Converged motif	No. of source motifs
AGAAGA	1484
AAGAAG	1375
GAAGAA	1225
others	12

RNAMotifModeler provides an option to predict sequence-structural consensus of different lengths. For short motifs, it is suggested to perform predictions starting from every potential motif candidate and generate a motif searching graph to inspect the global convergence. For longer motifs, however, generating such a graph will be computationally expensive. In this case, we conduct predictions starting from a sufficient number of motif candidates randomly picked from the motif space. The converged motif with the highest prediction power, measured by AUC, is selected as the optimal one.

2.3 Predicted sequence and structural features of SRSF1 binding regions

To better compare RNAMoifModeler predictions with the SRSF1 binding motif reported previously, here we focus on octamer predictions. Consistent with the sequence consensus predicted by MEME [4], the reference motif of SRSF1 identified using RNAMotifModeler is also 'AGAAGAAG'. The optimal parameters associated with the reference motif are displayed in Table 2. The first row listed the reference sequence motif identified while the following four rows include retained binding affinity ratios due to sequence alterations. The last row in Table 2, however, is constituted by unpaired probabilities for all nucleotides in the motif, indicating the optimal RNA secondary structure of SRSF1 binding regions. We note that every nucleotide of the predicted SRSF1 binding motif has a very high probability to be single-stranded, suggesting that SRSF1 proteins tend to bind on highly unpaired RNA regions.

Table 2. Predicted sequence-structural consensus of SRSF1

	A	G	A	A	G	A	A	G
A	1.00	0.17	1.00	1.00	0.24	1.00	1.00	0.81
G	0.79	1.00	0.65	0.90	1.00	0.84	1.00	1.00
C	0.52	0.32	0.50	0.16	0.35	0.02	0.34	0.63
U	0.75	0.15	0.39	0.63	0.09	0.06	0.73	0.55
UP	0.99	0.96	0.99	0.99	0.98	0.99	0.92	0.83

Based on the predicted optimal parameters, we obtained an AUC of 0.875 (Fig. 3 A) and an maximal accuracy of 0.803 (Fig. 3 B), which are both higher than the MEME-based prediction, of which the AUC is 0.86 and maximal accuracy is 0.78 [4].

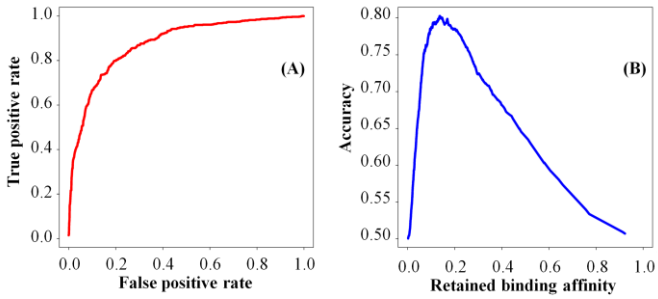

Fig. 3. ROC curve and accuracy curve describing the prediction power of RNAMotifModeler for SRSF1 proteins

To visualize the predicted SRSF1 sequence consensus more straightforwardly, positive gold standard sequence were scanned to search binding sites with binding affinities higher than the threshold 0.138, based on which a sequence logo (Fig. 4) was created by Weblogo [8]. This motif is consistent with the positional weight matrix (PWM) identified by MEME using the same gold standard sequences in our previous study [4], and is similar to the motifs found by other groups [9-11].

Fig. 4. Sequence consensus logo for SRSF1 proteins

2.4 SRSF1-RNA binding regions are significantly single-stranded

To further test the hypothesis that RNA regions bound by SRSF1 proteins are significantly unpaired, we compared

2904 binding sites predicted by RNAMotifModeler with a same number of controls binding sites, randomly selected in the same positive gold standard sequences. P-values were obtained from Wilcoxon rank sum tests on unpaired probabilities of nucleotides between predicted and randomly selected binding sites. All median unpaired probabilities of positive binding sites are significantly higher than controls (Fig. 5B). Wilcoxon tests were also performed on unpaired probabilities of nucleotides between predicted binding sites and random binding sites selected in negative gold standard sequences. For all the eight nucleotides, binding sites in positive gold standard sequences tend to be single-stranded (Fig. 5A).

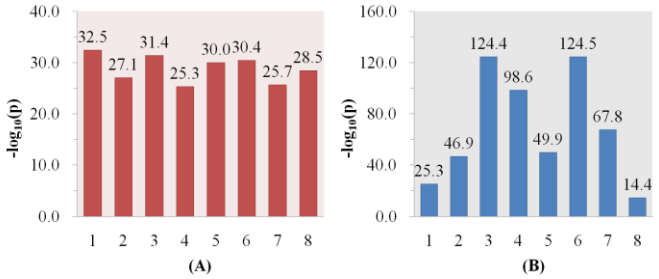

Fig. 5. P-values of nucleotides in the motif suggesting significant single-strandedness. The p-values are derived from Wilcoxon tests, with the alternative hypothesis that (A) predicted binding sites in positive gold standard sequences are more single-stranded than their counterparts in negative gold standard sequences, and (B) binding sites predicted by RNAMotifModeler are more single-stranded than randomly selected binding sites in positive gold standard sequences.

The two groups of Wilcoxon tests demonstrate that binding sites predicted by RNAMotifModeler are not only more single-stranded in positive gold standard sequences than negative controls, but also less structured than by chance within the same CLIP sequences. More interestingly, comparing Fig. 5 B with Fig. 4, we found that the second and fifth nucleotide of SRSF1 motif have much stronger sequence specificities but lower p-values of unpaired probabilities, while the third, fourth, sixth and seventh nucleotide are more significantly single-stranded but have less sequence specificities, suggesting that both the sequence and a lack of secondary structure may play complementary roles in SRSF1-RNA binding.

2.5 Predictions before and after incorporating RNA structure information

RNAMotifModeler can also predict consensus motifs without using structural information. Using the same positive and negative gold-standard sequences, we identified the same reference motif 'AGAAGAAG' and very similar retained binding affinity ratios due to sequence alterations. However, we obtained an optimal AUC of 0.853 and the maximal accuracy of 0.789, suggesting a slightly reduced prediction power when discarding RNA secondary structure information.

Using identified parameter matrix based on only sequences we predicted 2295 binding sites, of which 81% are commonly identified by incorporating RNA secondary structure information (Fig. 6 A). The unpaired probabilities of the other 437 binding sites are significantly lower than identified binding sites using both sequence and structural information (Fig. 6 B and 6 C). Except the third nucleotide of motif, all of the unpaired probabilities of these binding sites are even lower than background, indicating that binding sites predictions may result in a considerable number of false positives due to ignoring RNA secondary structures. Bringing in RNA secondary structure information, we found 1046 more binding sites. These binding sites may have low sequence specificities, but could be of high structure specificities. Although the AUC increases only by 0.023 after introducing RNA secondary structure information, false positive and false negative binding sites are both significantly reduced.

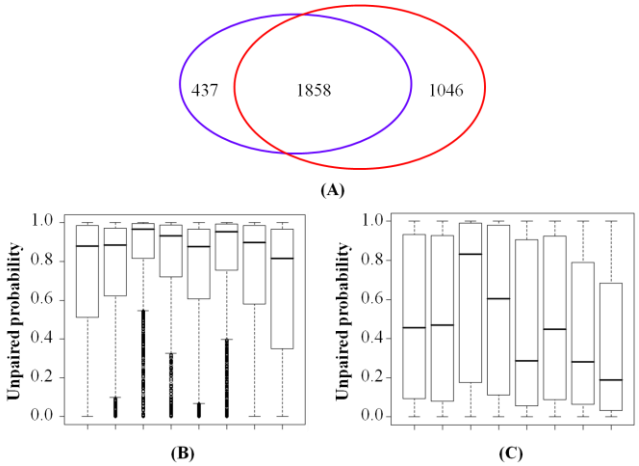

Fig. 6. Comparisons between predicted binding sites before and after incorporating RNA secondary structure information. (A) The number of binding sites predicted by RNAMotifModeler using only sequence information (blue ellipse) and after incorporating structure information (red ellipse); (B) Boxplots of unpaired probabilities of 1858 binding sites both predicted by the two methods; (C) Boxplots of unpaired probabilities of 437 binding sites only predicted without RNA secondary structure information

3 Discussions

In recent years, there is an increasing interest in using high-throughput sequencing technology to study protein-RNA binding specificities, but almost all of currently available bioinformatic approaches used for this purpose do not take into account RNA secondary structures, which have been demonstrated to have critical impact on protein-RNA binding in previous biochemical experiments. Thus, the motivation of our proposed model—RNAMotifModeler is to predict both structural and sequence specificities of protein-RNA binding regions.

RNAMotifModeler incorporates RNA secondary structure using RNAfold derived probabilities of nucleotides being paired with its neighbours. The preference for base-pairing probabilities over RNA secondary structures is due to a couple of concerns: a) It is very difficult to take into account RNA secondary structures directly in many real applications because of multiple RNA folding choices including optimal and sub-optimal structures; b) Unlike

MEMERIS, RNAMotifModeler tries to identify the optimal structural feature that is expected to represent the base pairing probability for each nucleotide in motif. Therefore, we did not use PU or EF scores [3], which are the measurements of single-strandedness of protein-binding regions in MEMERIS. c) The base-pairing probabilities predicted by RNAfold program [2] account for all possible secondary structures.

It is noted from our predictions that almost all unpaired probabilities of bases in the reference motif of SRSF1 predicted by RNAMotifModeler are close to 1, suggesting a very strong preference of SRSF1 to single-stranded RNA context. The statistical significance was further proved by two groups of Wilcoxon tests. These findings are consistent with previous evidences of SRSF1 proteins. It is known that SRSF1 protein contains an arginine-serine rich region (RS domain) and two RNA recognition motifs (RRMs), through which SRSF1 recognizes specific RNA regions [12, 13]. Importantly, RRM is one of the single-stranded RNA-binding domains of proteins [14]. Comparing the sequence consensus and p-values derived from Wilcoxon tests between the unpaired probabilities of predicted binding sites and negative controls, we propose that sequence and structural specificity may be two complementary factors that both facilitate the binding site recognition of SRSF1.

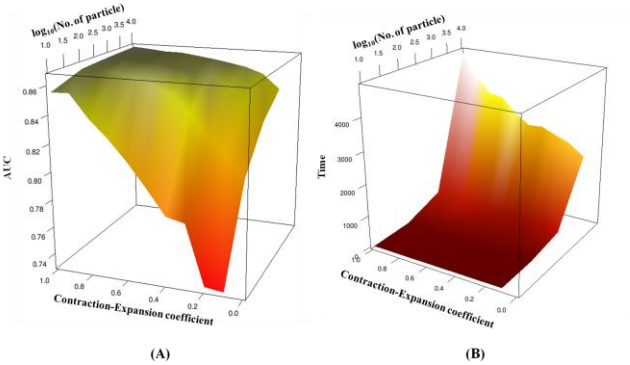

Fig. 7. 3D heatmaps illustrating the effects of the number of particles and the contraction-expansion coefficient in QPSO. (A) The prediction power measure by AUC, and (B) the time consumed are affected by the number of particles and the Contraction-Expansion coefficient, which are two critical parameters of QPSO.

RNAMotifModeler also provides an option to predict only sequence consensus motifs. This can be potentially applied to other fields that only focus on sequence specificities such as prediction of protein-DNA binding motifs. In the specific application to SRSF1, we found that the prediction power in this case is still comparable with MEME-based approach, although the AUC and maximum accuracy were both slightly reduced when RNA secondary structure information was not incorporated. Moreover, only using sequence specificity to predict binding sites could result in many false positives and false negatives.

Two parameters—the number of particles n_P and the contraction-expansion coefficient β of the Quantum Particle Swarm Optimization greatly affect the predicting accuracy of RNAMotifModeler. To estimate and set up these parameters prior to the optimization procedure, we did a series of hexamer motif searching tests with n_P enumerated from 10 to 10000 and β ranging from 0 to 1 for SRSF1 CLIP-seq data. The AUC scores resulted from optimizations using different combinations of these two parameters are presented in 3D heatmaps (Fig. 7A). We observed a much more rapid decrease in prediction power as β becomes lower when n_P is small. In contrast, when β is sufficiently high, the AUC score is not greatly affected by n_P. Thus, the greater n_P and β are, the higher prediction performance RNAMotifModeler can achieve. However, under the consideration of computational efficiency, we have to consider the time consumed in each test (Fig. 7B). The time consumed is exponential to the increment of the number of particles, and is not actually controlled by β. When n_P is 100 and β equals 1.0, RNAMotifModeler achieved a high AUC score of 0.86 within three minutes. These two parameters are then selected for all other optimizations for the SRSF1 dataset used in this study.

Convergence of optimization algorithms used in predicting protein-DNA or protein-RNA binding sites is a common concern due to a number of parameters needed to fit in model. In this report, we proposed a motif searching pathway and a motif searching graph to inspect whether or not the algorithm of RNAMotifModeler indeed has a good convergence regardless of the randomly initialized motif candidates. In the application to SRSF1 consensus motif, the convergence of randomly initialized motif candidates to final targets turned out to be very fast. Thus, for short motifs, we suggest generate such a motif searching graph in order to have a global overview of all possible converged motifs and their possible relationships.

Despite our successful characterization of the binding features of SRSF1 proteins, our future work will be applying RNAMotifModeler to studying specificities of other RNA binding proteins such as fox2, NOVA and EWS, for which high-throughput sequences are already available.

4 Methods

4.1 Predicting RNA base-pairing probabilities

One of the distinct features of RNAMotifModeler is that the information of secondary structures of the RNA regions bound by SRSF1 proteins is incorporated into the motif identification. For each nucleotide in the RNA fragment, we calculate the base pairing probability using the RNAfold function of the Vienna RNA package (version 1.8.5) [2]. The base pairing probability is used since it integrates likelihood of single-strandedness over multiple possible RNA secondary structures. For the CLIP-seq derived RNA fragments, these probabilities are generated based on the base pairing probability of base i being paired with base j, denoted as $p_{i,j}$. The binding probability of base i with all other neighbouring bases, defined as P_i, is calculated by:

$$P_i = \sum_{j=i+1}^{n_s} P_{ij} + \sum_{j=1}^{i-1} P_{ji} \quad (1)$$

where n_s is the length of sequence s. Similar strategies are also used elsewhere [15, 16].

4.2 Modelling protein-RNA binding affinities

In RNAMotifModeler, the consensus of each binding motif is defined by the following components: 1) the reference motif, a k-base RNA sequence on which the protein preferably binds; 2) retained binding affinity despite of a one-nucleotide deviation from reference motif to the sequence of one binding sites. For each k-base motif, there are $3k$ retained binding affinities that describe all the possible deviations from reference motif. For instance, if the i-th base of the reference motif and a specific binding site is m_i and f_i, respectively, the retained binding affinity is defined as μ_{i,m_i,f_i}; 3) a vector that denotes the optimal base pairing probability of k bases in the motif $\boldsymbol{\theta}=(\theta_i)$; and 4) the penalty for the deviation from the optimal base pairing probability α. All these parameters will be optimized iteratively. A matching score describing the similarity between an RNA fragment (F) and a reference motif (R) is defined:

$$S_{R,F} = \max_{l=1}^{L-k+1} \left(S_{R,F,l} \right) \tag{2}$$

where $S_{R,F,l}$ is the binding affinity for l-th binding site:

$$S_{R,F,l} = \prod_{i=1}^{k} \left(\left(\mu_{i,m_i,f_{l,i}} \right) \left(1 - \alpha \bullet \left| \theta_i - P_{f_{l,i}} \right| \right) \right) \tag{3}$$

where P_{f_i} represents the pairing probability of the i-th nucleotide in the RNA fragment F, calculated in Eq. (1). This matching score integrates the loss of binding affinity caused by both nucleotide and structure deviances from reference motif. We denote the parameter associated to the reference motif R as $\boldsymbol{\lambda}_R = (\boldsymbol{\mu}, \boldsymbol{\theta}, \alpha)_R$, where $\boldsymbol{\mu}$, $\boldsymbol{\theta}$ and α represent the $3k$ retained binding affinities, optimal base pairing probability of k bases, and the penalty for the deviation from the optimal base pairing probability, respectively.

4.3 Identify the optimal reference motif from CLIP-seq data

We adopted an iterative approach to identify the optimal reference motif and its associated parameters, using a Quantum Particle Swarm Optimization algorithm (QPSO) [17]. The iterative strategy includes the selection of reference motif R, and optimization of the parameters associated to the reference motif $\boldsymbol{\lambda}_R$. The overall procedure includes the following steps:

1. Motif initiation. Randomly select a motif candidate R_{init} from the motif searching space $\mathbf{M}=\{b_1b_2...b_k: b_1, b_2, ..., b_k \in \{A, G, C, U\}\}$ as the reference motif.

2. Parameter optimization. Optimize parameters associated with the reference motif by maximizing its ability for characterizing the CLIP-seq-derived RNA fragments.

Step 2.1. Parameter initiation. We first create n_P particles in the parameter space by randomly selecting numbers from U(0, 1).

Step 2.2. Particle evaluation. For each particle (parameters), we evaluate its capability for distinguishing the CLIP-seq-derived RNA fragment from background sequences. We plot an ROC (Receiver Operating Characteristic) curve by adjusting the matching score threshold, calculated in Eq. (2). The quality of the parameter will be evaluated based on the AUC (area under the curve) of the ROC plot.

Step 2.3. Particle update. Let $\boldsymbol{\lambda}_i^{selfbest}(t)$ and $\boldsymbol{\lambda}^{globalbest}(t)$ be the best individual particle i and the population of particles has met at the t-th iteration. As part of QPSO, each particle must converge to its local attractor $\boldsymbol{\lambda}_i^{pbest}$ [17]. Compute $\boldsymbol{\lambda}_i^{pbest}(t)$ and the mean of the best positions of all particles $\boldsymbol{\lambda}_i^{mbest}$ as follows:

$$\lambda_{i,j}^{pbest}(t) = (\varphi_1 \cdot \lambda_{i,j}^{selfbest}(t) + \varphi_2 \cdot \lambda_j^{globalbest}(t)) / (\varphi_1 + \varphi_2), \tag{4}$$

$$\lambda_j^{mbest}(t) = \sum_{i=1}^{n_P} \lambda_{i,j}^{pbest}(t) / n_P, \tag{5}$$

where φ_1 and φ_2 are random variables following U(0, 1);

QPSO employs Monte Carlo method to update parameters:

$$\lambda_{i,j}(t+1) = \begin{cases} \lambda_{i,j}^{pbest}(t) - \beta \cdot |\lambda_j^{mbest}(t) - \lambda_{i,j}(t)| \cdot \ln(1/u), & q \geq 0.5 \\ \lambda_{i,j}^{pbest}(t) + \beta \cdot |\lambda_j^{mbest}(t) - \lambda_{i,j}(t)| \cdot \ln(1/u), & q < 0.5 \end{cases}, \tag{6}$$

where β is called contraction-expansion coefficient controlling the convergence speed of QPSO; u and q are random variables which also follow U(0, 1).

Repeat Step 2 and Step 3 until $|\lambda^{globalbest}(t+1) - \lambda^{globalbest}(t)| < \varepsilon$ repeatedly, in which ε is a tolerance used here as the stop criterion;

3. Updating reference motifs. Based on the final parameter vector $\boldsymbol{\lambda}^{globalbest}$, the maximal binding affinity of motif candidate K in positive gold standard sequence F is:

$$a_{K,F} = \max_{\sigma \in \Omega_{K,F}} a_{K,F,\sigma}, \tag{7}$$

where $\Omega_{K,F}$ denotes the set of all binding sites for motif K in sequence F; $a_{K,F,\sigma}$ is also computed by Eq. (3).

In order to update the reference motif, from each positive fragment in the gold standard binding set, we selected the binding site that contributes to the positive selection (genomic loci with the highest binding affinity score). This potential binding site can be either the same as the reference motif, or different due to degeneracy. The reference motif will be further updated to the binding site that can represent largest amount of positive fragments in the gold standard binding set. Let n_F and n_M be the number positive gold standard sequences and the number of motif candidates, respectively. Let $S_{R_{init},F}$ be the maximal binding affinity computed using optimized parameters for the initial reference motif R_{init} in sequence F. To evaluate contribution

of each motif candidate, we define a motif contribution score matrix $\mathbf{c} = [c_{F,K}]_{F=1,2,...,n_S, K=1,2,...,n_M}$, in which

$$c_{F,K} = \begin{cases} 0, & a_{K,F} \neq S_{R_{init},F} \\ 1, & a_{K,F} = S_{R_{init},F} \end{cases}, \quad (8)$$

and a motif contribution score vector $\mathbf{v} = [v_K]_{K=1,2,...,n_M}$, in which:

$$v_K = \sum_{F=1}^{n_S} c_{F,K} . \quad (9)$$

We denote the motif associated with the maximum score in **v** as R_{max}. If $R_{max}=R_{init}$, meaning the initialized reference motif accounts for the most contribution to the retained binding affinities, then we stop the iteration; otherwise, let R_{max} be the next R_{init}, and repeat step 2 and 3 until convergence.

4.4 RBP binding motif logo

RNAMotifModeler provides a parameter matrix consisting of retained binding affinity ratios due to sequence mutations and structure alterations at each base. For the ease of visualization, we provide a method to generate a Positional Weight Matrix (PWM). Once *RNAMotifModeler* reaches a convergence, a set of optimal parameters and reference motif will be acquired, as well as a cutoff score of binding affinity at the peak of the accuracy curve. We trace back subsequently to each positive gold standard sequence to identify binding sites with binding affinities higher than the cutoff score. Finally, using these positive binding sites, we calculate the PWM and create a corresponding logo based on the *Weblogo* tool [8].

ACKNOWLEDGMENT

This work is supported by the grant from National Institutes of Health, R21AA017941 (to YL), R01GM085121 (to JRS), and the Indiana Genomics Initiative of Indiana University (supported in part by the Lilly Endowment, Inc.).

5 References

[1] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," *FEBS letters,* vol. 582, pp. 1977-1986, 2008.

[2] I. L. Hofacker, "RNA secondary structure analysis using the Vienna RNA package," in *Curr Protoc Bioinformatics*, 2008/04/23 ed. vol. Chapter 12, 2004, pp. Unit 12 2.

[3] M. Hiller, R. Pudimat, A. Busch, and R. Backofen, "Using RNA secondary structures to guide sequence motif finding towards single-stranded regions," *Nucleic Acids Research,* vol. 34, pp. e117, 2006.

[4] J. R. Sanford, X. Wang, M. Mort, et al., "Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts," *Genome Res,* vol. 19, pp. 381-94, Mar 2009.

[5] E. Buratti, A. F. Muro, M. Giombi, D. Gherbassi, A. Iaconcig, and F. E. Baralle, "RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon," *Molecular and cellular biology,* vol. 24, pp. 1387, 2004.

[6] J. R. Sanford, J. Ellis, and J. F. Caceres, "Multiple roles of arginine/serine-rich splicing factors in RNA processing," *Biochemical Society Transactions,* vol. 33, pp. 443-446, 2005.

[7] R. Karni, E. De Stanchina, S. W. Lowe, R. Sinha, D. Mu, and A. R. Krainer, "The gene encoding the splicing factor SF2/ASF is a proto-oncogene," *Nature Structural & Molecular Biology,* vol. 14, pp. 185-193, 2007.

[8] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res,* vol. 14, pp. 1188-90, Jun 2004.

[9] M. Caputi, G. Casari, S. Guenzi, R. Tagliabue, A. Sidoli, C. A. Melo, and F. E. Baralle, "A novel bipartite splicing enhancer modulates the differential processing of the human fibronectin EDA exon," *Nucleic Acids Res,* vol. 22, pp. 1018-22, Mar 25 1994.

[10] J. Ramchatesingh, A. M. Zahler, K. M. Neugebauer, M. B. Roth, and T. A. Cooper, "A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer," *Mol Cell Biol,* vol. 15, pp. 4898-907, Sep 1995.

[11] R. Tacke and J. L. Manley, "The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities," *EMBO J,* vol. 14, pp. 3540-51, Jul 17 1995.

[12] J. C. K. Ngo, K. Giang, S. Chakrabarti, et al., "A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1," *Molecular cell,* vol. 29, pp. 563-576, 2008.

[13] J. C. Hagopian, C. T. Ma, B. R. Meade, C. P. Albuquerque, J. C. K. Ngo, G. Ghosh, P. A. Jennings, X. D. Fu, and J. A. Adams, "Adaptable molecular interactions guide phosphorylation of the SR protein ASF/SF2 by SRPK1," *Journal of molecular biology,* vol. 382, pp. 894-909, 2008.

[14] S. D. Auweter, F. C. Oberstrass, and F. H. T. Allain, "Sequence-specific binding of single-stranded RNA: is there a code for recognition?," *Nucleic Acids Research,* vol. 34, pp. 4943, 2006.

[15] M. Hiller, R. Pudimat, A. Busch, and R. Backofen, "Using RNA secondary structures to guide sequence motif finding towards single-stranded regions," *Nucleic Acids Res,* vol. 34, pp. e117, 2006.

[16] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers,* vol. 29, pp. 1105-19, May-Jun 1990.

[17] J. Sun, B. Feng, and W. Xu, "Particle swarm optimization with particles having quantum behavior," in *Evolutionary Computation, 2004. CEC2004. Congress on*, 2004, pp. 325-331.