Bioinformatics, 32, 2016, i351–i359
doi: 10.1093/bioinformatics/btw259
ISMB 2016

OXFORD

RCK: accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data

Yaron Orenstein[1], Yuhao Wang[1] and Bonnie Berger[1,2,*]

[1]Computer Science and Artificial Intelligence Laboratory and [2]Math Department, MIT, Cambridge, MA, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Protein–RNA interactions, which play vital roles in many processes, are mediated through both RNA sequence and structure. CLIP-based methods, which measure protein–RNA binding *in vivo*, suffer from experimental noise and systematic biases, whereas *in vitro* experiments capture a clearer signal of protein RNA-binding. Among them, RNAcompete provides binding affinities of a specific protein to more than 240 000 unstructured RNA probes in one experiment. The computational challenge is to infer RNA structure- and sequence-based binding models from these data. The state-of-the-art in sequence models, Deepbind, does not model structural preferences. RNAcontext models both sequence and structure preferences, but is outperformed by GraphProt. Unfortunately, GraphProt cannot detect structural preferences from RNAcompete data due to the unstructured nature of the data, as noted by its developers, nor can it be tractably run on the full RNACompete dataset.

Results: We develop RCK, an efficient, scalable algorithm that infers both sequence and structure preferences based on a new *k*-mer based model. Remarkably, even though RNAcompete data is designed to be unstructured, RCK can still learn structural preferences from it. RCK significantly outperforms both RNAcontext and Deepbind in *in vitro* binding prediction for 244 RNAcompete experiments. Moreover, RCK is also faster and uses less memory, which enables scalability. While currently on par with existing methods in *in vivo* binding prediction on a small scale test, we demonstrate that RCK will increasingly benefit from experimentally measured RNA structure profiles as compared to computationally predicted ones. By running RCK on the entire RNAcompete dataset, we generate and provide as a resource a set of protein–RNA structure-based models on an unprecedented scale.

Availability and Implementation: Software and models are freely available at http://rck.csail.mit.edu/

Contact: bab@mit.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein–RNA interactions play vital roles in many processes in the living cell (Rinn and Ule, 2014). These include modulation and effect of a wide variety of cellular processes, including RNA replication, repair, recombination and post-transcriptional regulation (Rinn and Ule, 2014). More than 1500 genes in the human genome are thought to code for RNA-binding proteins (RBPs), making this family one of the largest families in the human proteome (Gerstberger *et al.*, 2014). Most RBPs bind RNA through both sequence and structure. Thus, better characterization of RBP sequence- and structure-specific binding preferences can improve our understanding of post-transcriptional gene regulation.

RNA structure is commonly considered at the level of secondary structure (Washietl *et al.*, 2012). RNA secondary structure is represented by base-pairing of nucleotides, and can be efficiently predicted from its sequence when only tree-like structures are allowed. A single RNA molecule may fold into different conformations, termed its 'ensemble' of structures, where the most likely one is the minimum free energy structure (Steffen *et al.*, 2006). These ensembles can be represented as either different combinatorial structures

(represented as graphs by RNAshapes (Janssen and Giegerich, 2014)) or average probabilities over the ensemble (represented as position-specific probability vectors by RNAplfold (Lorenz *et al.*, 2011)). New experimental methods for protein–RNA binding, such as icSHAPE, can probe RNA structure both *in vitro* and *in vivo* (Spitale *et al.*, 2015). In many cases, the exact base-pairing is not as informative as the structural context: whether the binding site is in a loop, external region or base-paired (Leontis *et al.*, 2006). A common assumption is that most proteins prefer to bind accessible sites (i.e. those that are unlikely to be base-paired), and a few prefer a specific structural context (Li *et al.*, 2014; Ray *et al.*, 2009).

New experimental high-throughput (HTP) techniques have been developed to uncover protein–RNA interactions on a genome-wide scale at single-nucleotide resolution. For example, HITS-CLIP, CLIP-seq and RIP-seq measure protein–RNA interactions *in vivo* in a HTP manner (König *et al.*, 2012). However, protein RNA-binding is influenced by a variety of factors, such as other RBPs, which either compete for the same binding site or co-bind as a complex, and RNA structure. In addition, experimental output is significantly affected by technological artifacts and noise (Fu and Ares, 2014; Kishore *et al.*, 2011). While the end goal is to understand and predict *in vivo* binding, *in vitro* experiments currently have higher resolution and less noise. Moreover, as accurate experimental measurements of RNA structure are scarce, methods rely on computational prediction of RNA structure, which is more accurate *in vitro* than *in vivo* (Rouskin *et al.*, 2014). Thus, *in vitro* experiments provide valuable complementary information that may enable us to learn intrinsic protein RNA-binding preferences.

Towards this aim, high-throughput *in vitro* methods have been developed to study the binding preferences of RBPs (Lambert *et al.*, 2014; Ray *et al.*, 2009). A prime example is RNAcompete (Ray *et al.*, 2009), in which a specific protein binds to a set of pre-designed oligos and binding is measured by hybridization to complementary probes on a microarray. The experimental output includes binding intensities to more than 240 000 probe sequences. These sequences, divided uniformly into Set A and Set B, together cover each RNA 9-mer at least 16 times. The first study of 9 experiments included sets of sequences from some structured RNA probes (Ray *et al.*, 2009). In the second study the probes were designed as unstructured to be accessible for the protein to bind (Ray *et al.*, 2013). The latter study reported the binding of more than 200 human RBPs and provided the first-of-its-scale dataset of protein–RNA binding measurements (Ray *et al.*, 2013). However, only sequence-based binding models were inferred from those due to the probes' unstructured design.

Several methods have been developed to infer sequence- and structure-based models for protein–RNA binding. MEMERIS was the pioneer method (Hiller *et al.*, 2006) to explore binding sites in unpaired regions by extending the well-known MEME (Bailey *et al.*, 2015) to incorporate unpaired nucleotide probabilities. A more recently developed method, RNAcontext, infers a PWM model and structural context preferences from RNAcompete and CLIP data (Kazan *et al.*, 2010). RNA structure is represented as probabilities of structural contexts (paired, hairpin loop, multi loop, inner loop and external), predicted by RNAplfold (Lorenz *et al.*, 2011). The state-of-the-art in discovering structure- and sequence-based binding preferences is GraphProt (Maticzka *et al.*, 2014). The secondary structure of each RNA sequence is represented as combinatorial graphs. Up to three most-probable structures are predicted per sequence using RNAshapes (Janssen and Giegerich, 2014). Graph-based support vectors (Costa and De Grave, 2010) are used to train a model, where the features are local subgraphs (representing base-paired and adjacent nucleotides) and hyper-graphs (representing structural contexts).

Unfortunately, both GraphProt and RNAcontext have clear limitations in predicting protein–RNA binding. The prediction of highly probable combinatorial structures, as in GraphProt, while saving time, may not reveal the full ensemble of possible structures; the aggregation of less likely structures has been shown to be more informative of RNA-binding than minimum free energy structures (Li *et al.*, 2014). On the other hand, while RNAcontext uses structural context probabilities to represent the complete ensemble of possible structures, its sequence model is a position weight matrix (PWM), which is inherently position independent; every position in the model contributes independently to binding. Thus, PWMs cannot model dependencies between binding site positions, as opposed to GraphProt, where subgraph features encode sets of nucleotides with their structural relationships.

To resolve these issues, we present a new algorithm RCK (short for RNAcontext-k-mer) (see Fig. 1). We follow the intuition that RNA-binding preferences require more complex models, incorporating both sequence and structure. The algorithm extends RNAcontext by a *k*-mer sequence- and structure-based binding model. In particular, RCK uses the same input and optimization procedure as RNAcontext to infer model parameters, but in a *k*-mer based context that better captures local preferences, and a more efficient implementation. We applied our algorithm to RNAcompete data and inferred, for the first time, structure-based models from this large-scale data of protein–RNA binding, as previous methods were not applied due to design limitations based on the unstructured nature of the data. We show that our method significantly outperforms both the sequence-based and structure-based state-of-the-art methods for *in vitro* binding prediction. It is also much more efficient, both in running time and memory usage, which is what enables scalability to larger datasets and more complex binding models. While in our limited tests RCK is on par with extant methods in predicting *in vivo* binding based on predicted RNA structure, we demonstrate that RCK can easily incorporate experimentally measured RNA structure to improve *in vivo* binding prediction.

2 Methods

2.1 RCK *k*-mer model for protein–RNA binding

We extended the model of RNAcontext (Kazan *et al.*, 2010) to account for *k*-mer sequence and structure preferences. First, we define a *structural probability vector* for sequence s to be a vector of length $|s|$ where each element is a distribution over the possible structural contexts (e.g. paired and unpaired) in that position. The model assumes that the binding intensity to *k*-mer w with structural probability vector p is a multiplication of its sequence and structure preferences, modeled by Θ:

$$N(w,p,\Theta) = N^{\text{seq}}(w,\Theta) \times C(w,p,\Theta) \quad (1)$$

where $N^{\text{seq}}(w,\Theta)$ is the sequence binding score, given by:

$$N^{\text{seq}}(w,\Theta) = \sigma(b_s + \phi_w) \quad (2)$$

in which ϕ_w is the sequence *k*-mer score of w and b_s is the sequence bias.

$C(w,p,\Theta)$ is the structure binding score, given by:

$$C(w,p,\Theta) = \sigma(b_p + \sum_{a\in A} \Gamma_{w,a} \times \sum_{i=1}^{k} p_{a,i}) \quad (3)$$

where $\Gamma_{w,a}$ is the structural preference of *k*-mer w to structural context a, $p_{a,i}$ is the probability of position i in w to be in structural

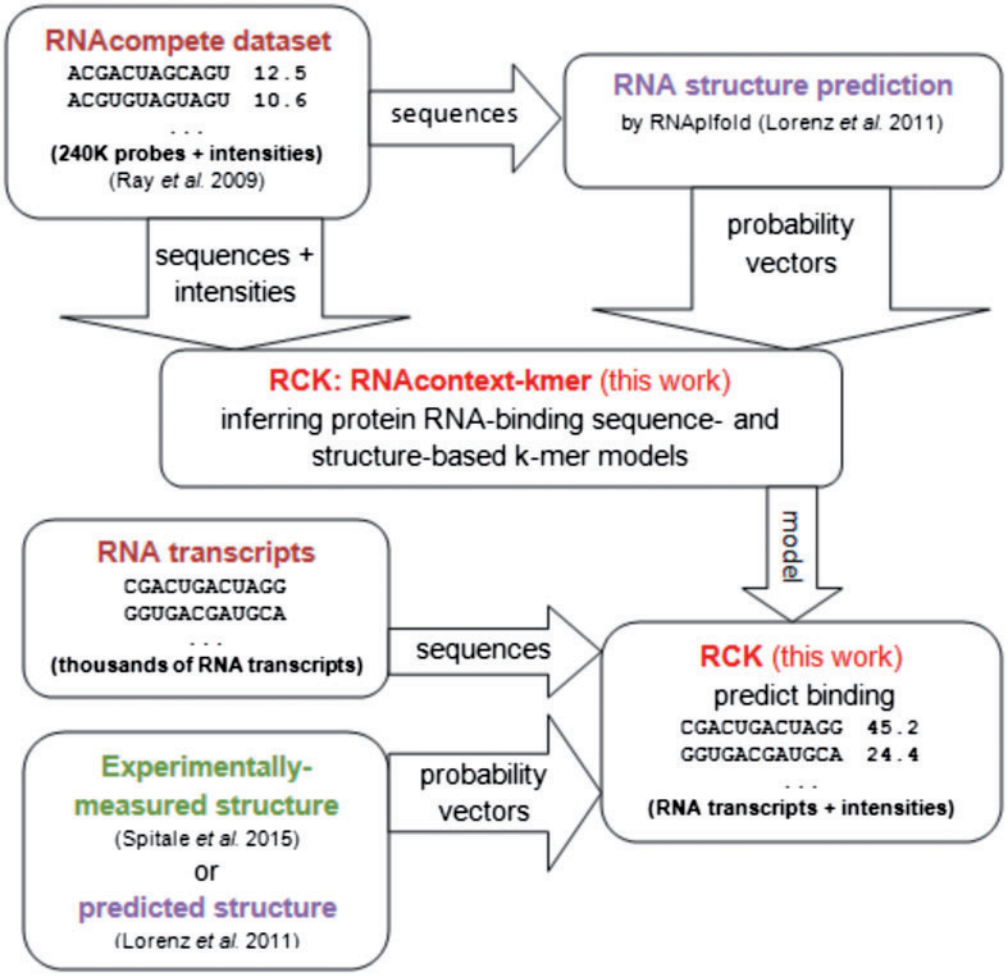

Fig. 1. A flowchart depicting the input, output and function performed by RCK. To infer a new binding model, RCK receives as input an RNAcompete dataset (Ray *et al.*, 2009), containing thousands of sequences and corresponding binding intensities, together with predicted structure probabilities of the probe sequences (by RNAplfold (Lorenz *et al.*, 2011)). RCK learns a sequence- and structure-based binding model, which can be later used to predict binding intensities of new RNA sequences, for which structure has been experimentally measured (Spitale *et al.*, 2015) or computationally predicted

context a, and b_p is the structure bias. In both (2) and (3), σ is the logistic function: $\sigma(x) = (1 + \exp(-x))^{-1}$. The bias terms, b_s and b_p, are meant to scale the sequence and structure preferences. For example, if structure does not play a significant role in binding, b_p may take a large positive value.

The binding score of sequence S with structural probability vector P is:

$$f(S, P, \Theta) = 1 - \prod_{t=0}^{|S|-k} 1 - N(S_{t+1:t+k}, P_{t+1:t+k}, \Theta) \qquad (4)$$

where $S_{t+1:t+k}$ and $P_{t+1:t+k}$ are the k-long sub-vectors starting at index $t+1$ of S and P, respectively. The assumption is that every position in sequence S has an independent probability of being bound. The probability of sequence S not being bound, is the product of not binding at any position.

2.2 Parameter inference

The model parameters are inferred to reduce the error in probe intensity prediction. For N sequences $S = \{s^1, \ldots, s^N\}$ with structure probability vectors $P = \{p^1, \ldots, p^N\}$, measured binding intensities $R = \{r^1, \ldots, r^N\}$ and predicted intensities $\widehat{R} = \{\hat{r}^1, \ldots, \hat{r}^N\}$, where $\hat{r}^i = \alpha \cdot f(s^i, p^i, \Theta) + \beta$, we optimize the parameters Θ to minimize the sum of squared errors:

$$E(\Theta, \alpha, \beta) = \sum_{i=1}^{N} (r^i - \hat{r}^i)^2 + \delta \sum_{\phi_w \in \Theta} \sigma(b_s + \phi_w) \qquad (5)$$

The least squares cost function $E(\Theta, \alpha, \beta)$ is minimized using the L-BFGS method (Byrd *et al.*, 1995). Note that we modified the regularization term to suit the new model. Regularizing by the sum of k-mer sequence scores is intended to have few k-mers with high scores and most k-mers with close-to-zero scores.

2.3 RNA secondary structure prediction

RNA secondary structural context profiles were predicted using a variant of RNAplfold (Lorenz *et al.*, 2011). In this variant, probabilities for four structural contexts are calculated per position: hairpin loop, inner loop, multi loop and external region (Kazan *et al.*, 2010). The probability for a position being paired is assigned, so that the total sum is 1. These probabilities, represented as 5 vectors, each the length of the sequence, are provided together with the sequences as input to RCK and RNAcontext.

2.4 Evaluating model performance

To evaluate the new k-mer model, we used the original dataset on which RNAcontext was developed (Kazan *et al.*, 2010). This dataset included nine paired experiments, each pair testing a different protein (Ray *et al.*, 2009). For each pair, we trained the model on experiment A and tested on experiment B. We gauged the performance by the Pearson correlation of measured and predicted binding intensities. To be able to see the performance gain irrespective of model width (i.e. k), we ran RCK and RNAcontext on width 5.

2.5 Running time evaluation

We measured RNAcontext, GraphProt and RCK training runtimes and memory usage on the RNCMPT00001:setA dataset. We gauged the user running time, excluding structure prediction times. Time and memory usage were measured using the `/usr/bin/time` command in Unix. User time and maximum resident set size are reported. Running times and memory usage were benchmarked on a single CPU of a 20-CPU Intel Xeon E5-2650 (2.3 GHz) machine with 384 GB 2133 MHz RAM.

We ran GraphProt in the following way: `perl GraphProt.pl -mode regression -fasta <sequence file> -affinities <intensities file>`. RCK and RNAcontext were run with default parameters and model width 5.

2.6 *In vitro* binding prediction evaluation

To evaluate the performance of algorithms on *in vitro* binding prediction we used the RNAcompete dataset (Ray *et al.*, 2013). The set includes 244 experiments, each containing binding intensities to more than 240 000 sequences. The sequence set was designed as a union of two sets, A and B, such that each has similar k-mer coverage. For each experiment, we trained a model on sequences from set A and predicted the intensities on set B. Performance was determined by the Pearson correlation of predicted and measured intensities of set B. Outlier intensities were clamped as done in the Deepbind study (Alipanahi *et al.*, 2015): all intensities above the 0.5 percentile were clamped to the value of the 0.5 percentile.

Five methods were compared in this evaluation. Sequence-based methods included: PWMs from Ray *et al.* (Ray *et al.*, 2013), MatrixREDUCE (Foat *et al.*, 2006) and Deepbind (Alipanahi *et al.*, 2015). Results were taken from (Alipanahi *et al.*, 2015). Structure-based methods included RCK and RNAcontext. For both, model width was optimized by a simple 2-fold cross-validation on the training set. Width ranges of RNAcontext and RCK were 4–12 and 4–6, respectively.

2.7 *In vivo* binding prediction evaluation

To evaluate the performance of algorithms on *in vivo* binding prediction we used the dataset curated in the GraphProt study (Maticzka *et al.*, 2014). 24 CLIP experiments were collected, each providing a set of *in vivo* binding sites. For them, control sequences were extracted from unbound regions of the same bound transcripts. Both binding sites and control sequences were flanked by 150nt on both ends. The overlap with RNAcompete *in vitro* data included 23 pairs of CLIP and RNAcompete experiments, covering 10 proteins (Supplementary Table S3). We trained a model on a complete RNAcompete experiment and tested its prediction accuracy in ranking binding sites higher than control sequences. We reported the average AUC over the pairs of experiments for each protein.

RCK and RNAcontext model width was optimized in a conventional 2-fold cross-validation. Width ranges were 4–6 and 4–12, respectively. To account for different sequence lengths, scores for each sequence were assigned by averaging over the k-mer scores in that sequence. RNAplfold was run on the sequences and its flanks, while only the sequence and its predicted probabilities were used for testing. Deepbind was run with the '–average' option to average k-mer scores over a sequence. Its pre-computed models were publicly available (Alipanahi *et al.*, 2015).

2.8 Structural source evaluation

To evaluate the effect of different RNA structure sources, we used CLIP and icSHAPE data (Spitale *et al.*, 2015). Probability vectors of experimentally-measured RNA structure and CLIP-seq data were downloaded from the GEO database (accession numbers GSE60034 and GSE64168, respectively). Binding site peaks were extracted as in the original study (Spitale *et al.*, 2015) using a 40nt window size. We selected peaks that had measured structure both *in vivo* and *in vitro* over all nucleotides, summing up to 4102 positive sequences. As a control, we randomly selected 4102 40nt-long sequences that had measured RNA structure. For computational structure prediction, we flanked binding sites and control sequences by 150nt on each end, which were only used for structure prediction by RNAplfold (Lorenz *et al.*, 2011) and later discarded for the testing (as done in (Maticzka *et al.*, 2014)).

2.9 Model visualization

To visualize the model in an interpretable way, we collapsed it into a PWM and structure parameters in the following way. For each pair of k-mer and structure context (w, a), we calculated its score by:

$$M(w, a, \Theta) = \sigma(b_s + \phi_w) \times \sigma(b_p + \Gamma_{w,a}) \quad (6)$$

We defined the (w, a) with the highest score as the consensus. To derive a PWM, we used all k-mers at Hamming distance 1 from w, and used their score in the structural context a as the weight in the position of difference. For structure preferences, we used k-mer w scores in different structural contexts as the weights. Sequence logos were plotted using motifStack (Developer J and Developer L, 2015).

3 Results

3.1 RCK: a new algorithm to infer protein–RNA binding preferences

We developed a new model for sequence and structure protein–RNA binding preferences (Fig. 2A). The model is an extension of the model used by the RNAcontext algorithm (Kazan *et al.*, 2010), which is based on a PWM and a vector of structure preferences. Our new model is more complex, both in sequence and in structure. Sequence-wise, each k-mer is assigned a unique sequence score. Structure-wise, each k-mer, where the optimal value of k is typically five for RNAcompete data (Supplementary Fig. S2), is assigned a vector of structural preferences. See Section 2 for details and formal definitions.

To demonstrate the improvement achieved by these model changes, we ran RNAcontext on the original RNAcompete dataset on which it was developed (Ray *et al.*, 2009). This benchmark includes nine pairs of RNAcompete experiments. The sequences were designed so that together they cover all 9-mers in unstructured regions and all 7-mers in hairpin loops. We inferred a model from experiment A and tested it on experiment B as in the original study (Kazan *et al.*, 2010). Performance was gauged by Pearson correlation of predicted and measured intensities.

Results show that the new RCK model performs significantly better than the old model. First, we demonstrate the benefit of using

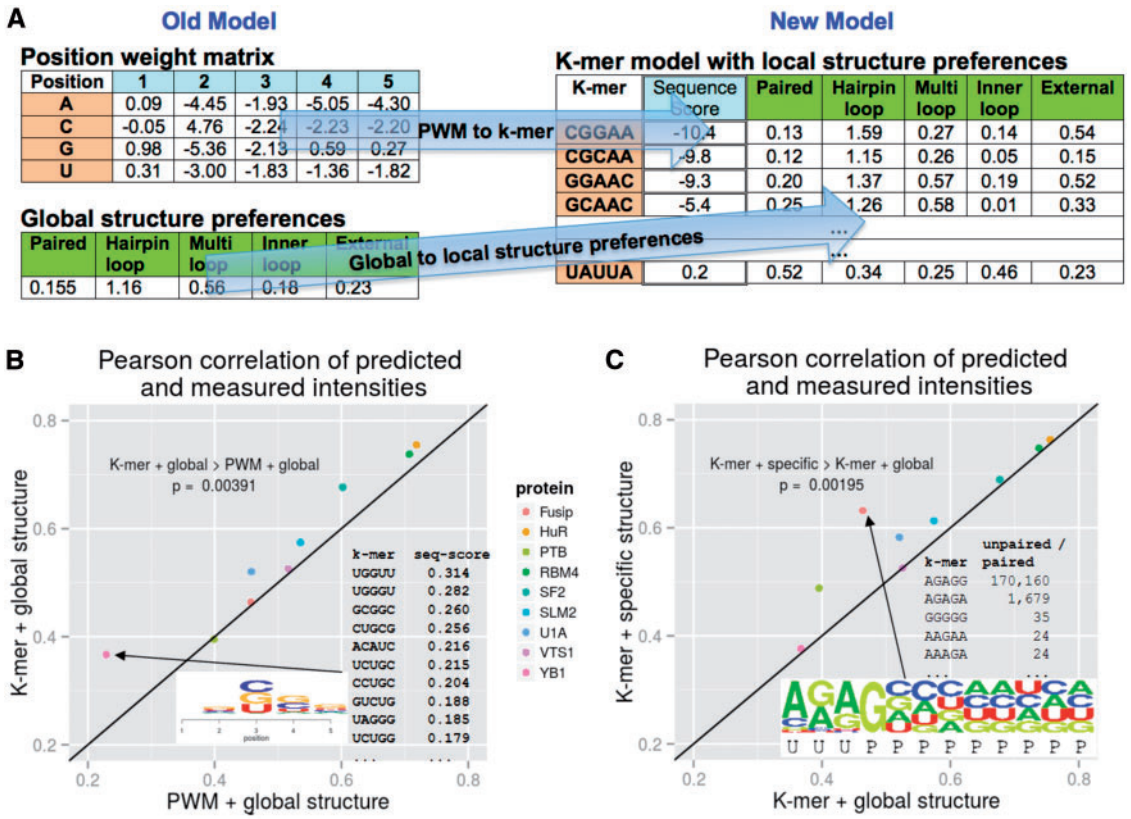

Fig. 2. A new *k*-mer based model for protein–RNA binding (RCK). (**A**) The new *k*-mer model with *k*-mer specific sequence and structure preferences. Each *k*-mer has a sequence score and a vector of structural preferences. In the old model, *k*-mers were assigned scores according to a position weight matrix and global structural preferences. (**B**) The benefit of newly using sequence *k*-mer scores. For nine pairs of experiments, a model was trained on set A of the pair and tested on set B using RNAcontext and different sequence and structure models. Pearson correlation was used to evaluate prediction accuracy. Average Person correlation improved from 0.514 to 0.557. YB1 prefers to bind to several distinct *k*-mers, which are not modeled well by a single PWM. (**C**) The benefit of using *k*-mer specific structural preferences (on top of *k*-mer sequence scores). Average Person correlation improved from 0.557 to 0.602. The same datasets were used as in (B). In Fusip's model, the *k*-mer with the highest ratio of binding in unpaired compared to paired context is AGAGG, as was also observed by alignment of binding sites (Kazan *et al.*, 2010)

k-mer sequence scores (Fig. 2B). Pearson correlation is higher by more than 0.01 for 5 out of the 9 proteins, and on average by 0.043 (P-value$= 3.91 \times 10^{-3}$, Wilcoxon rank-sum test). The improvement is particularly high for protein YB1. According to our results (Fig. 2B), YB1 prefers to bind GU-rich *k*-mers on top of the known literature motifs CAUC and CACC (Wei *et al.*, 2012; Wu *et al.*, 2015). Secondly, we observe the benefit of having specific structure preferences for each *k*-mer. The average Pearson correlation increased by 0.045 as compared to the *k*-mer model with global structure preferences (P-value$= 1.95 \times 10^{-3}$), and by more than 0.01 in 5 out of the 9 proteins. For Fusip1, which regulates splicing, the improvement is remarkable (0.17 in Pearson correlation). In the original RNAcontext study, alignment of predicted binding sites revealed that Fusip1 prefers to bind an unpaired region next to a paired region (Fig. 2C). Notably, such a hybrid preference cannot be modeled by global structure preferences, which assign the same structural preference for all *k*-mers. For complete results, see Supplementary Table S1.

3.2 RCK is much faster and memory-efficient than the state of the art

We compared the running times and memory usage of the different algorithms to infer sequence and structure protein–RNA binding preferences: RCK, RNAcontext and GraphProt. For this task, we ran each algorithm on RNAcompete experiment RNCMPT00001: setA. We used default parameters and excluded the running time of the parameter optimization (see Section 2 for details). In addition, we excluded the running time of the structure prediction for two reasons. First, it is negligible compared to the runtime of the algorithm. RCK and RNAcontext use RNAplfold (Lorenz *et al.*, 2011), which takes less than 20 min on this sequence set. GraphProt uses RNAshapes (Steffen *et al.*, 2006), which takes less than 10 min. In addition, since each RNAcompete experiment is performed on the same set of sequences, the structure prediction needs to be run only once.

Results show that RCK is slightly faster than RNAcontext and much faster than GraphProt in model training (Supplementary Fig. S1). For a complete run on RNCMPT00001:setA, RCK and RNAcontext terminated in less than 2 h (101 and 107 min, respectively), while GraphProt required more than 7 days (10 660 min). Despite the fact that RCK trains a model that is exponentially greater in size than RNAcontext's model, it is slightly faster thanks to our improved implementation of the L-BFGS optimization. In our implementation, the change in the parameters is calculated by one pass over the data, as opposed to the original implementation where a pass was made for each parameter. This reduces each iteration's runtime to be linear in the size of the input, instead of the size of the input times the size of the model, as done in the original implementation. RCK and RNAcontext are much faster than GraphProt, which is based on support vector regression to learn the binding models.

In terms of memory usage, RCK is more efficient than both RNAcontext and GraphProt (Supplementary Fig. S1). The memory demand of GraphProt on RNCMPT00001:setA is more than 12GB, while RCK and RNAcontext require 700 MB and 2 GB, respectively. RCK is more efficient than RNAcontext thanks to our improved implementation that allocates less memory for data and parameters, despite the increased model size.

3.3 RCK is more accurate in *in vitro* binding prediction than the state of the art

To gauge the performance of RCK compared to extant methods, we used the comprehensive dataset of RNAcompete, which includes 244 experiments (Ray *et al.*, 2013). We clamped outlier intensities of each experiment as described in the Deepbind study (Alipanahi *et al.*, 2015), and trained a model on set A sequences. Performance was gauged by Pearson correlation of predicted and measured set B probe intensities. For sequence-based methods, we used the results published in the Deepbind study, comparing algorithms Ray et al. (Ray *et al.*, 2013), MatrixREDUCE (Foat *et al.*, 2006) and Deepbind (Alipanahi *et al.*, 2015). We added to the comparison the results of RNAcontext. We did not include GraphProt in the comparison as, in addition to its inability to infer structural preference from RNAcompete data due to the unstructured nature of the data (Maticzka *et al.*, 2014), it is memory- and time-intensive; it requires more than 7 days and 12GB of memory to run on Set A of one RNAcompete experiment. For complete details, see Section 2.

RCK significantly outperformed all methods in *in vitro* binding prediction (Fig. 3A). When comparing RCK to sequence-based methods, it outperformed the state-of-the-art Deepbind, which achieved an average AUC of 0.409 as compared to 0.460 for RCK (P-value= 3.67×10^{-11}, Wilcoxon rank-sum test) (Fig. 3B). RCK outperformed the structure-based method RNAcontext as well (Fig. 3C), which achieved an average AUC of 0.433 (P-value= 2.36×10^{-17}). Notably, RNAcontext did not perform significantly better than Deepbind (P-value = 0.093). For complete results, see Supplementary Table S2.

3.4 RCK is as accurate at *in vivo* binding prediction as the state of the art

To gauge the performance of RCK compared to extant methods on *in vivo* binding prediction, we used the dataset curated as part of the GraphProt study (Maticzka *et al.*, 2014). The dataset includes 24 CLIP experiments, each containing thousands of experimentally validated *in vivo* binding sites and control sequences extracted from

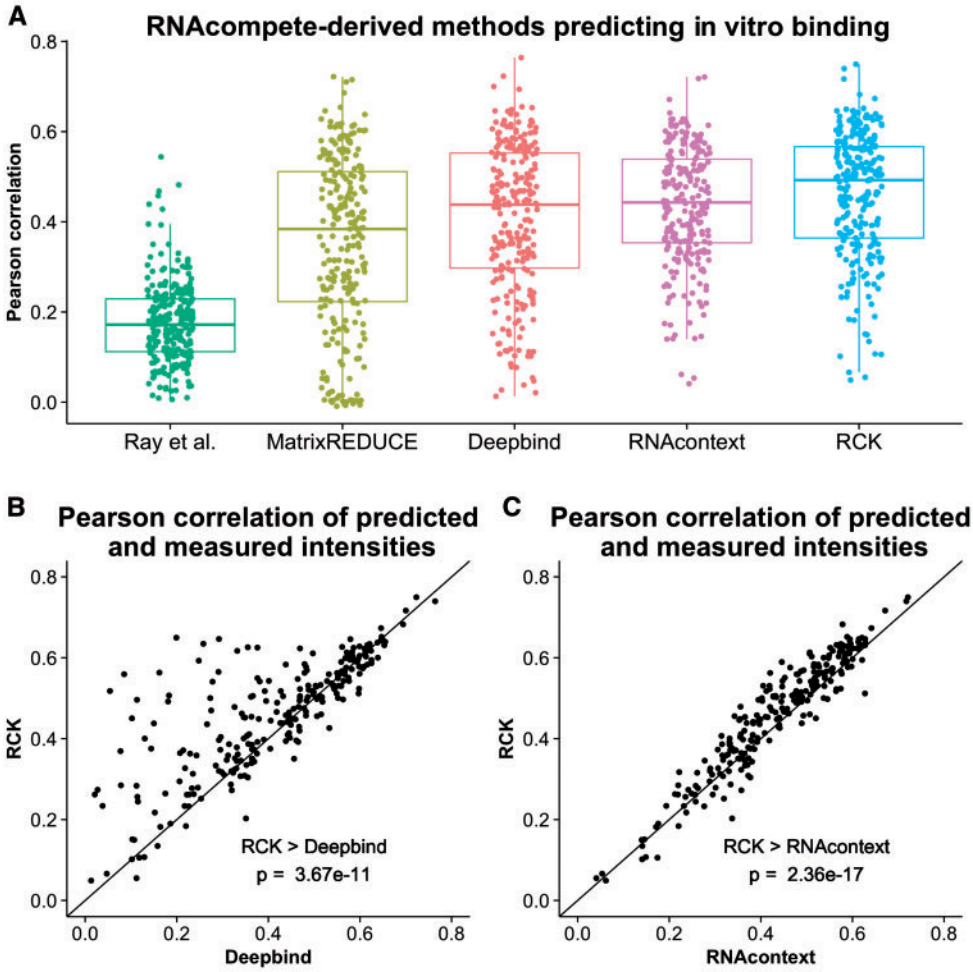

Fig. 3. Performance in predicting *in vitro* binding. For each RNAcompete experiment, a model was trained on Set A sequences and tested on Set B. Results are over 244 experiments. Performance gauged by Pearson correlation of predicted and measured intensities. (**A**) Boxplots of correlations for different methods. RNAcontext and RCK utilize RNA secondary structure. (**B**, **C**) Dot-plot comparison of RCK to Deepbind and RNAcontext, respectively. *P*-values calculated by Wilcox rank-sum test

unbound regions of the same transcripts. Each binding site and control sequence is flanked by adjacent 150nt, both downstream and upstream, to allow for accurate structure prediction (using RNAplfold (Lorenz *et al.*, 2011)). The overlap with RNAcompete's dataset covers 10 proteins from 21 RNAcompete and 12 CLIP experiments (Ray *et al.*, 2013). We could not use the curated *in vivo* dataset of the RNAcompete study (Ray *et al.*, 2013), as the sequences in it did not include flanks, and thus structure prediction would be inaccurate. For the sequence-based method, we used the state-of-the-art Deepbind (Alipanahi *et al.*, 2015), and for structure-based, RNAcontext. We trained RCK and RNAcontext models on a complete set of sequences of each RNAcompete experiment. Deepbind models were publicly available. We reported the performance in predicting *in vivo* binding by average AUC over the pairs of RNAcompete and CLIP experiments for each protein. See Section 2 for details.

Results show that the performance of all methods is comparable. On 10 proteins, no method outperforms the other two on the majority of proteins (Fig. 4A). RCK performs the best (albeit not significantly), achieving a median AUC of 0.803, compared to 0.791 and 0.778 for Deepbind and RNAcontext, respectively (Fig. 4B). In a pairwise comparison between RCK, Deepbind and RNAcontext, no method is significantly better (P-values >0.279, Wilcoxon rank-sum test). Two reasons may hamper the accuracy of *in vitro* models in predicting *in vivo* binding. First, *in vivo* data is known to be noisy and suffer from experimental biases (Fu and Ares Jr, 2014; Kishore *et al.*, 2011). Moreover, RNA structure prediction is less accurate *in vivo* than *in vitro* (Rouskin *et al.*, 2014). At this stage, more datasets with higher quality are needed in the overlap between CLIP and RNAcompete to derive more definitive conclusions. For complete results, see Supplementary Table S3.

3.5 RCK can easily be applied to experimentally measured RNA structure probabilities

One of the benefits of RCK is its ability to incorporate experimentally measured RNA structure probabilities. RCK receives as input probability vectors of different structural contexts of the input sequences in single nucleotide resolution. These can be either computationally predicted (e.g. by RNAplfold (Lorenz *et al.*, 2011)) or experimentally measured (e.g. by icSHAPE (Spitale *et al.*, 2015)). To demonstrate the effect of using experimental probabilities, we used available CLIP and icSHAPE experiments, performed on the same cells that also had an RNAcompete experiment on the same protein (see Section 2 for details). Unfortunately, only the HuR protein, which had five RNAcompete experiments, was found to overlap. We compared the effect of five different structural sources: *in vivo* icSHAPE, *in vitro* icSHAPE, *in silico* (RNAplfold), no probabilities (sequence scores only) and uniform probabilities. Since icSHAPE reports only unpaired probabilities, we trained a model based on two structural contexts: paired and unpaired. Performance was measured by AUC in predicting HuR binding sites.

Results show that RCK can benefit from experimentally measured RNA structure in predicting *in vivo* binding (Fig. 5). icSHAPE *in vivo* measurements are more accurate than *in vitro* measurement in four out of five experiments, and more accurate than predicted structure in four out of five experiments. When using uniform structure probabilities or sequence scores alone, performance decreased substantially. This can likely be explained by HuR's strong preference to bind unpaired regions (Li *et al.*, 2010), which was automatically encoded into RCK's models. We note that additional experimental measurements of RNA structure and protein–RNA binding on the same cells are needed to evaluate the benefit of experimentally measured RNA structure for the task of *in vivo* binding prediction. For complete results, see Supplementary Table S4.

4 Discussion

In this study, we developed RCK, a new algorithm to infer complex models of protein–RNA binding. RCK uses a k-mer based model for sequence preferences and specific structural context preferences for each k-mer based on probability profiles. We demonstrated the accuracy of our algorithm on the most comprehensive set of *in vitro* data, where its inferred models were significantly more accurate in predicting *in vitro* binding than the state-of-the-art. Moreover, we showed that RCK is capable of incorporating either predicted or

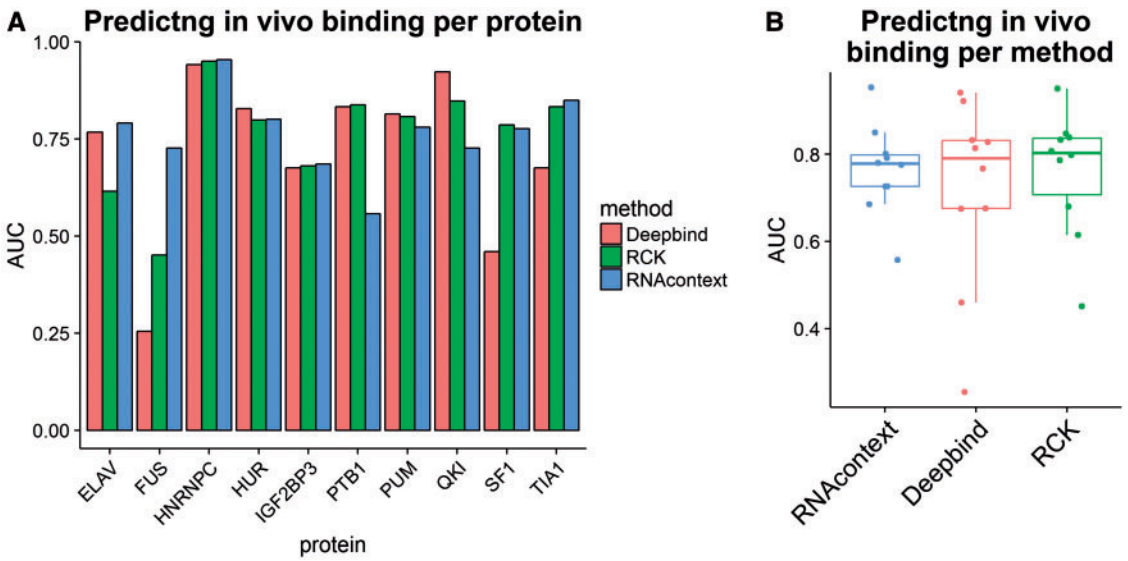

Fig. 4. Performance in predicting *in vivo* binding. For each pair of RNAcompete and CLIP experiments on the same protein, a model was trained on the former and tested on the latter. 23 pairs overlap with the GraphProt study and RNAcompete dataset, covering 10 proteins in 21 RNAcompete and 12 CLIP experiments. Performance per protein is gauged by average AUC. (**A**) Bar-plot of average AUCs of different methods per protein; RNAcontext and RCK utilize RNA secondary structure. (**B**) Boxplots of methods' performance

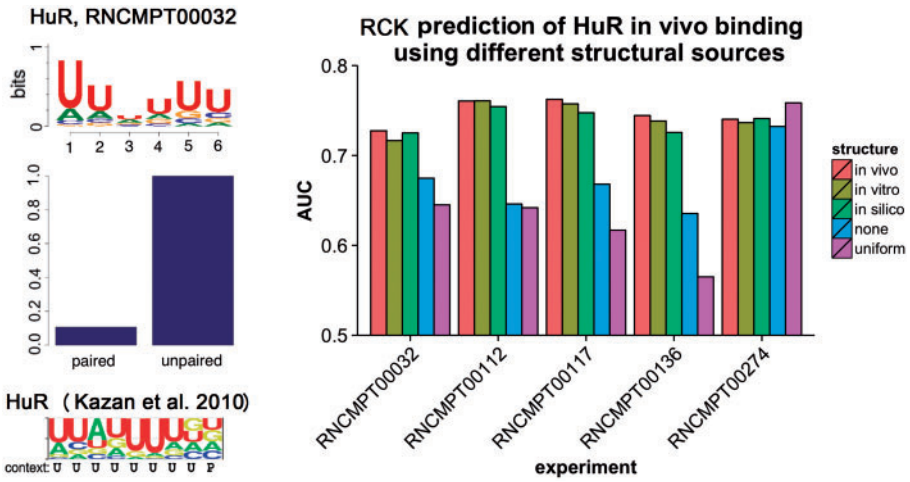

Fig. 5. Structural sources effect on RCK in vivo binding prediction. HuR binding model was learned from different RNAcompete experiments. As an example, the sequence and structural preferences learned from RNCMPT00032 are shown (left upper and middle, respectively), as well as the sequence logo and structural context from (Kazan *et al.*, 2010). RCK models were used to predict *in vivo* binding based on structural probabilities derived from different sources (right): *in vivo* (icSHAPE (Spitale *et al.*, 2015)), *in vitro* (icSHAPE), *in silico* (RNAplfold (Lorenz *et al.*, 2011)), no probabilities (sequence scores only) and uniform probabilities. Performance of HuR binding prediction was gauged by AUC

experimentally measured RNA structure probabilities to improve *in vivo* binding prediction. Unfortunately, *in vivo* datasets that overlap RNAcompete are too few to reach definitive conclusions regarding performance.

The success of our more complex model shows that protein–RNA binding is better modeled using a more complex k-mer based model. The position-independence assumption that is inherent in the position weight matrix has been challenged for a long time (Eggeling *et al.*, 2014, 2015). To make no assumptions, we used a list of all possible k-mers and a score for each one. For structure preferences, individual k-mers may have different structure preferences; global preferences common to all k-mers may be too limiting. By incorporating unique preferences for each k-mer, we allow k-mers to have unique structural preferences.

The k-mer model is the most comprehensive binding model, but it also has its disadvantages. The number of parameters increases significantly compared to simple models, increasing the risk of over-fitting the training data. For example, the original model included $4k$ sequence parameters (where k is the width of the model) and Σ_A structure parameters (where Σ_A is the number of structural contexts). The k-mer model has 4^k sequence parameters, and $4^k \times \Sigma_A$ structure parameters. When comparing different k values, $3 \leq k \leq 7$ (see Supplementary Fig. S2), we observed a decrease in performance for $k \geq 6$, which we explain by over-fitting to the training set. For example, when $k = 7$ and $\Sigma_A = 5$ the model contains 98 298 parameters (see below), while the training set has only around 120 000 samples. To avoid over-fitting, we used a 2-fold cross-validation on the training set to find the optimal width in the range $4 \leq k \leq 6$. Another disadvantage of the k-mer model lies in its visualization. As the model is more complex, it is more difficult to visualize the complete landscape of sequence and structure preferences in an interpretable manner.

On the implementation side, we were able to extend RNAcontext and make it more efficient both for running time and memory usage. In our improved implementation, each iteration of the optimization procedure is linear in the size of the input and number of parameters, as compared to the size of the input times the number of parameters in the original implementation (which would have been very costly with our expanded model of $4^k \times (\Sigma_A + 1)$ parameters). Moreover, L-BFGS, the optimization procedure used in RCK, runs in time $O(N + D)$, where D is the size of the input and N the number of parameters (Liu and Nocedal, 1989). SVR, used by GraphProt, requires $O(D^2N)$ time (Burges, 1998), which is infeasible for large datasets, as in RNAcompete.

We see several potential extensions to this work given the new model and inferred sequence and structure preferences. First, as we expanded the model to the k-mer model, we limited it to relatively small k's ($k \leq 6$) due to the size of the datasets. It may be worthwhile to utilize machine learning algorithms with stricter regularization terms that infer sparse models, such as Lasso (Tibshirani, 2011). In this way, k may be increased without jeopardizing the accuracy of the models. Second, the current model assumes a single structural context for the whole k-mer. Structural preferences may be position-specific in single-nucleotide resolution. Expanding the model to position-specific structure preferences may be possible by either using a simpler sequence model (e.g. PWM) or learning a sparse model (as mentioned above), as imposing such an extension on the k-mer model increases the number of parameters by a factor of k, the width of the model. Third, the application of *in vitro* protein–RNA binding models to predict *in vivo* data is still lacking. We believe that by using experimentally measured RNA structures we can improve *in vivo* binding prediction, as we demonstrated here on a small scale. In this aspect, more CLIP and experimentally measured RNA structure datasets are needed as well as improvements in their quality.

To summarize, we developed RCK, a new algorithm to infer protein–RNA binding preferences. RCK is highly accurate at predicting *in vitro* binding. By applying it to the RNAcompete dataset, we were able to newly uncover the structural preferences of more than 200 proteins, which we make available as a resource on the RCK website. We hope that the new algorithm and its inferred models will provide a rich platform and resource for future studies to better understand the binding mechanism underlying and regulatory roles of protein–RNA binding.

Acknowledgements

We thank Ryan A. Flynn for helping with the icSHAPE and CLIP-seq analysis of the published data from their lab.

Funding

This work was supported by the National Institutes of Health [NIH grant R01GM081871].

Conflict of Interest: none declared.

References

Alipanahi,B. *et al*. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol*., **33**, 831–838.

Bailey,T.L. *et al*. (2015) The MEME suite. *Nucleic Acids Res*., **43**, W39–W49.

Burges,C.J. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov*., **2**, 121–167.

Byrd,R.H. *et al*. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput*., **16**, 1190–1208.

Costa,F. and De Grave,K. (2010). Fast neighborhood subgraph pairwise distance kernel. In: *Proceedings of the 26th International Conference on Machine Learning, Omnipress*, pp. 255–262.

Developer L. and Developer L. (2015) motifStack: Plot stacked logos for single or multiple DNA, RNA and amino acid sequence. R package version 1.14.0. http://bioconductor.org/packages/release/bioc/html/motifStack.html.

Eggeling,R. *et al*. (2014) On the value of intra-motif dependencies of human insulator protein CTCF. *PloS One*, **9**, e85629.

Eggeling,R. *et al*. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 375.

Foat,B.C. *et al*. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.

Fu,X.D. and Ares Jr,M. (2014) Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet*., **15**, 689–701.

Gerstberger,S. *et al*. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet*., **15**, 829–845.

Hiller,M. *et al*. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res*., **34**, e117–e117.

Janssen,S. and Giegerich,R. (2014) The RNA shapes studio. *Bioinformatics*, **31**, 423–425.

Kazan,H. *et al*. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol*., **6**, e1000832.

Kishore,S. *et al*. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.

König,J. *et al*. (2012) Protein–RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet*., **13**, 77–83.

Lambert,N. *et al*. (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell*, **54**, 887–900.

Leontis,N.B. *et al*. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol*., **16**, 279–287.

Li,X. *et al*. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096–1107.

Li,X. *et al*. (2014) Finding the target sites of RNA-binding proteins. *Wiley Interdisc. Rev.: RNA*, **5**, 111–130.

Liu,D.C. and Nocedal,J. (1989) On the limited memory BFGS method for large scale optimization. *Math. Program*., **45**, 503–528.

Lorenz,R. *et al*. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol*., **6**, 1.

Maticzka,D. *et al*. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*., **15**, R17.

Ray,D. *et al*. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol*., **27**, 667–670.

Ray,D. *et al*. (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.

Rinn,J.L. and Ule,J. (2014) Oming in on RNA–protein interactions. *Genome Biol*., **15**, 10–1186.

Rouskin,S. *et al*. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.

Spitale,R.C. *et al*. (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519**, 486–490.

Steffen,P. *et al*. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.

Tibshirani,R. (2011) Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)*, **73**, 273–282.

Washietl,S. *et al*. (2012) Computational analysis of noncoding RNAs. *Wiley Interdisc. Rev.: RNA*, **3**, 759–778.

Wei,W.J. *et al*. (2012) YB-1 binds to CAUC motifs and stimulates exon inclusion by enhancing the recruitment of U2AF to weak polypyrimidine tracts. *Nucleic Acids Res*., **40**, 8622–8636.

Wu,S.L. *et al*. (2015) Genome-wide analysis of YB-1-RNA interactions reveals a novel role of YB-1 in miRNA processing in glioblastoma multiforme. *Nucleic Acids Res*., **43**, 8516–8528.