Deep Learning Generative Architecture for Deriving RNA-Binding Protein (RBP) Binding Motifs and Applications to Multiple Types of RBP Sequencing Data

Matthew Agar-Johnson (andrewid: magarjoh)

May 12, 2019

Introduction

RNA-Binding Protein (RBP) interactions with RNA are significantly involved in a variety of gene regulatory processes, including post-transcriptional modification, gene splicing, and localization. There are a variety of molecular biology methods for isolating RBP-bound RNAs for sequencing, including cross-linking immunoprecipitation (CLIP-seq), and more protein-specific methods such as Translating Ribosome Affinity Purification (TRAP-seq) [1]. RNA-Bind-n-Seq [2] is a technique for sequencing RBP-bound RNA and quantifying binding affinity. Different concentrations of purified RBP are added to a cell-free RNA sample, which can subsequently be immunoprecipitated. One of the drawbacks of CLIP-Seq is that it is difficult to distinguish sequences that bind to the lone RBP, or sequences (which may have different identifying motifs) that bind to RBP complexes. Sites that bind to protein complexes tend to have more limited roles in transcriptional regulation than do motifs which selectively bind to the lone protein [2].

Computational approaches for predicting RNA-Binding motifs have traditionally relied on classical machine learning approaches, notably MEME, which uses expectation maximization to fit a mixture model and discover sequence motifs. A modified version of this algorithm, MEMERIS, uses a similar technique to determine RNA structure and single-stranded regions [3]. These techniques suffer from the feature engineering problem inherent to all classical machine learning problems. Only recently, deep learning techniques, such as convolutional neural nets (CNNs) and recurrent neural nets (RNNs), have been used to develop structure-prediction methods from sequencing data alone [4]. While these techniques are becomingly increasingly robust, they mostly rely solely on RNA sequence. In the past year, methods have been developed to create deep learning models that incorporate both RNA sequence and secondary structure [5], which can also be derived computationally [6]. By incorporating structural elements and motifs into RNA-binding protein analysis, insights on specific RBP binding preferences can be achieved [7]. The goal of this project was to build on existing deep-learning generative architectures in the literature and determine if different structural motifs were present in

Bind-n-Seq and CLIP-Seq datasets for the same RNA Binding Protein. Multiple datasets of each type were used and examined for intra- and inter-consistency.

Materials and Methods

Data All datasets were examining the RNA-Binding Protein IGF2BP2 (insulin-like growth-factor 2 mRNA-binding protein 2) and were obtained from the ENCODE project website. 2 Independent CLIP-Seq Experiments were used (ENCFF307AUO, ENCFF643OXD). CLIP-Seq experiments were performed on the human K562 chronic myelogenous leukemia (CML) lymphoblast cell line. Bind-n-Seq data was taken from ENCODE as well, both from cell-free samples with RBP concentrations of 20nM (ENCFF998ZLV) and 320nM (ENCFF833VXY). The input library to the Bind-n-Seq experiment (ENCFF148TFJ) was used as a negative control. In principle, lower concentrations of purified protein should pull down motifs with stronger affinity, but overall consistency between Bind-n-Seq samples is expected. For all samples, two separate datasets of size $n = 100,000$ were used, one as a training set and another as a test set to measure reconstruction error of the model.

Model Architecture The problem of RBP motif discovery requires a generative model that can be trained by unsupervised learning. For these purposes, an undirected probabilistic graphical model called a Restricted Boltzmann Machine (RBM) was used [8]. An RBM encodes a joint probability distribution between hidden units h and visible units v with parameters θ as follows:

$$p(v, h; \theta) = \frac{1}{Z(\theta)}\, \sigma(E(v, h; \theta))$$

where $Z(\theta)$ is the partition function, σ is the logistic function, and E is an energy function. The energy function is defined as:

$$E(v, h; \theta) = -\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} v_i h_j - \sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j$$

Where W is a matrix of pairwise parameters for v and h, and b, c are bias vectors for v and h respectively.

Encoding RNA Primary Sequence Zhang et al. [7] explored the use of 'topic models' originally developed for NLP in field of RNA motif discovery. A topic model finds latent, hidden 'topics' given a series of documents $\mathcal{D}$, and a bag of words contained in a 'dictionary'. This project follows the model outlined in Zhang et al., and defines an RNA 'document' as a single sequencing read, and a k-mer motif candidate as an RNA 'word'. The size of the target motifs k is decided before experimentation, and the 'dictionary' $\mathcal{K}$ will consist of all possible k-mers.

Sequencing reads d are encoded into binary 'topic vectors' of length 4^k (All possibilities of 4-letter RNA alphabet) where $v_i = 1$ if $\mathcal{K}_i \subseteq d$, and 0 otherwise. The 'topic vectors' are treated as the visible units in the bottom level RBM, as shown below in Figure 1.

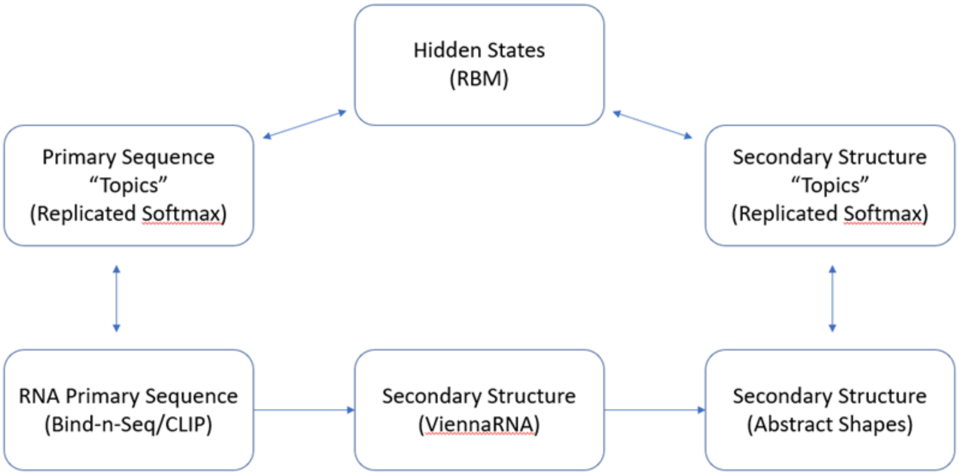
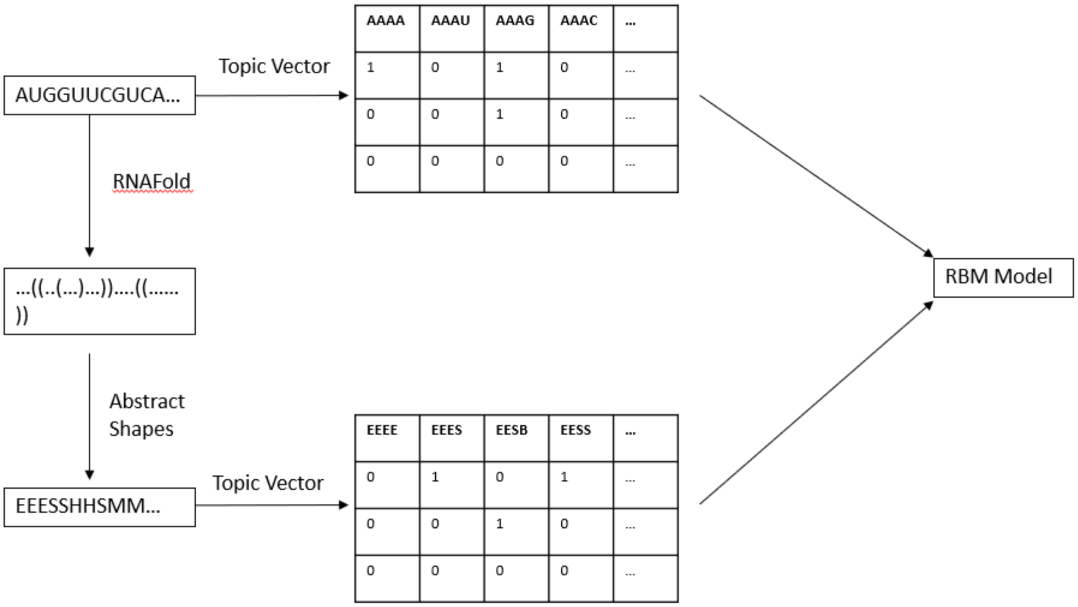


Figure 1: Overall deep learning architecture used for RBP motif discovery. Primary sequence reads are encoded as binary vectors using the topic model described below. Raw reads are then passed through ViennaRNA to determine likely secondary structures based on free energy. The dot-bracket notation is converted into an abstract shape representation, which can then be encoded as a binary topic vector like the primary sequence. Both of these are passed into individual RBMs, the hidden features of which serve as visible units for the top-layer RBM.

Encoding RNA Secondary Structures First, RNA secondary structures for a given set of sequencing reads were determined by the available ViennaRNA 2.0 Package RNAFold algorithm [9]. The algorithm returns reads in dot-bracket notation, where '.' represents an unpaired base, and '(' and ')' represent two bases paired together. This notation is not suitable for the topic model described above, so the dot-bracket notated outputs of RNAFold were further interpreted into 'abstract shapes'. This is a modified version of the procedure of Zhang et al. [7], which relies on a 5-letter alphabet 'EBSHM', where 'E' represents an unpaired base external to a loop structure, 'B' represents a bulge, 'S' a stem component of a stem-loop, 'H' representing an unpaired base in a hairpin loop, and 'M' representing a multiloop. Once dot-bracket notation was converted into the above described abstract shapes notation, it could be encoded into topic vectors in the same way as the primary sequences. It is important to note that due to the nature of the abstract shapes depiction, certain combinations are not possible (e.g. 'EEEH' would denote four unpaired bases in a row, three of which are arbitrarily classified as external and one of which is in a non-existent loop). These motifs were eliminated in a post-processing step. Due to the larger size of the secondary structure alphabet, and to limited available computing capacity, the size of the secondary structure topic vectors was restricted to all the non-zero (i.e. possible/occurring) motifs or to the limit of 4^k, with k being the length of primary sequence motifs. In practice, there were few secondary motifs that were identifiable in the reads, so this was not an issue.



AAAA	AAAU	AAAG	AAAC	...
1	0	1	0	...
0	0	1	0	...
0	0	0	0	...

EEEE	EEES	EESB	EESS	...
0	1	0	1	...
0	0	1	0	...
0	0	0	0	...

Figure 2: Data encoding for RBP motif discovery. Reads are entered into RNAFold and subsequently converted into the abstract shapes alphabet. Given n reads, $N \times K$ topic vector matrices are constructed, which will be treated as visible units for the RBM.

RBM Training To train the RBM, the Contrastive Divergence (CD) algorithm is used [10], with learning rate $\alpha = .01$. Two RBMs with separate parameters are trained on primary and secondary topic vectors. The hidden variables of each of these models were selected to be $h = 500$ for each, and were concatenated to form the visible unit vector for the top-layer RBM. This is an important step given the obvious fact that primary RNA sequence and structural features are not independent. The top-layer was set to have 100 hidden features, which were later used to generate motifs. After each epoch of training, the mean-squared-error of reconstruction on both the training and test datasets. The model was deemed to have converged when

$$err_{primary,t} + err_{secondary,t} < \epsilon(err_{primary,t-1} + err_{secondary,t-1})$$

For $\epsilon = .001$. This generally occured in under 50 epochs of training.

RBP Motif Generation To generate RBP sequence and structural motifs from the trained model, the variables h_{top} were initialized to random, and vectors p_{prim} and p_{sec}

were sampled according to the model. p_{prim} is a vector of probabilities of RNA motifs such that $p_{prim} = p(\mathcal{K}_i)$. Motifs that are enriched, i.e. above background probability, are sampled at each step. The criterion for 'enrichment' is defined by hyperparameters α, β indicating the number of standard deviations above the mean probability to be classified as enriched. These parameters were adjusted to keep the outputs in manageable size for WebLogo, which was used to visualize the resulting generated outputs. Following sampling, the two probability vectors p_{prim} and p_{sec} were used to sample a new vector h_{top} and the process was repeated to generate new samples, such that only the first iteration was truly random and the rest were sampled from the model itself.

Results

Reconstruction Error Training and test set reconstruction error was calculated at each epoch of training and plotted below.

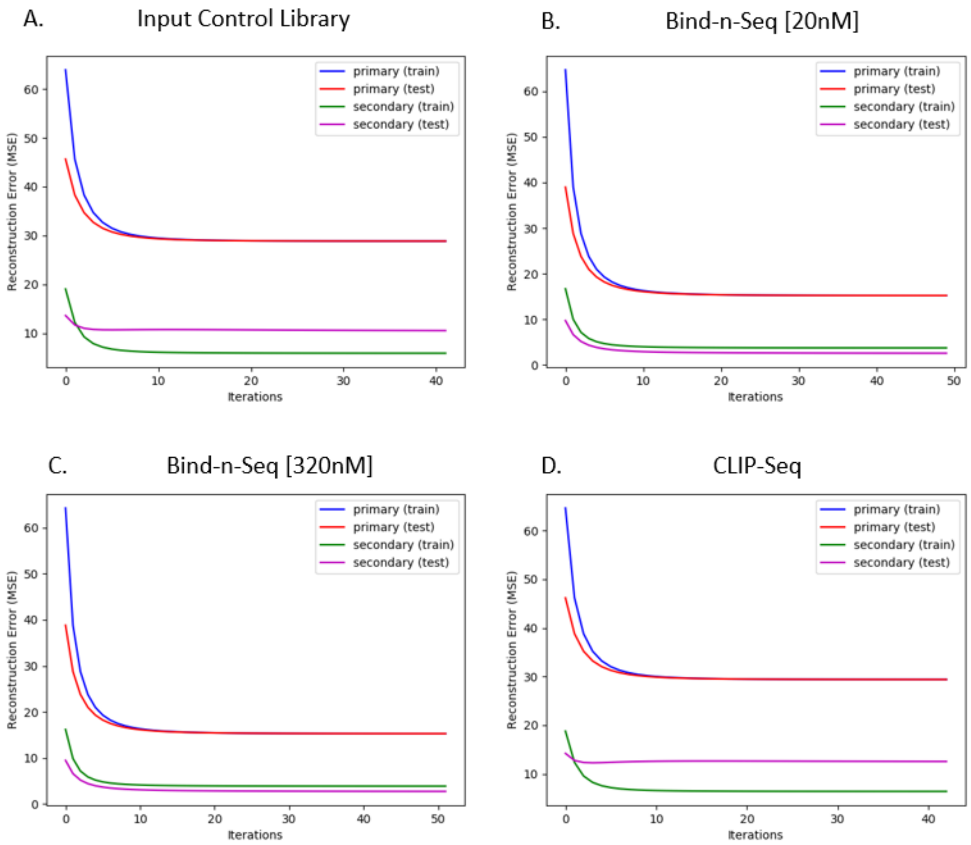


Figure 3: Reconstruction error per training iteration for (A) Bind-n-Seq input library control, (B) Bind-n-Seq 20nM RBP concentration, (C) Bind-n-Seq 320nM RBP concentration, (D) CLIP-Seq (averaged between 2 datasets).

It is noted that the reconstruction error for the secondary motifs on the test set in the Bind-n-Seq experiments is notably lower than the CLIP, which appears to be comparable to the input control, which has no IP whatsoever. All models seem to show some improvement through training, but on both primary and secondary motifs, improvement in the Bind-n-Seq is significantly larger.

Structural Motif Generation 10,000 secondary structural motifs were sampled from the models and passed into WebLogo to visualize any significant motifs.

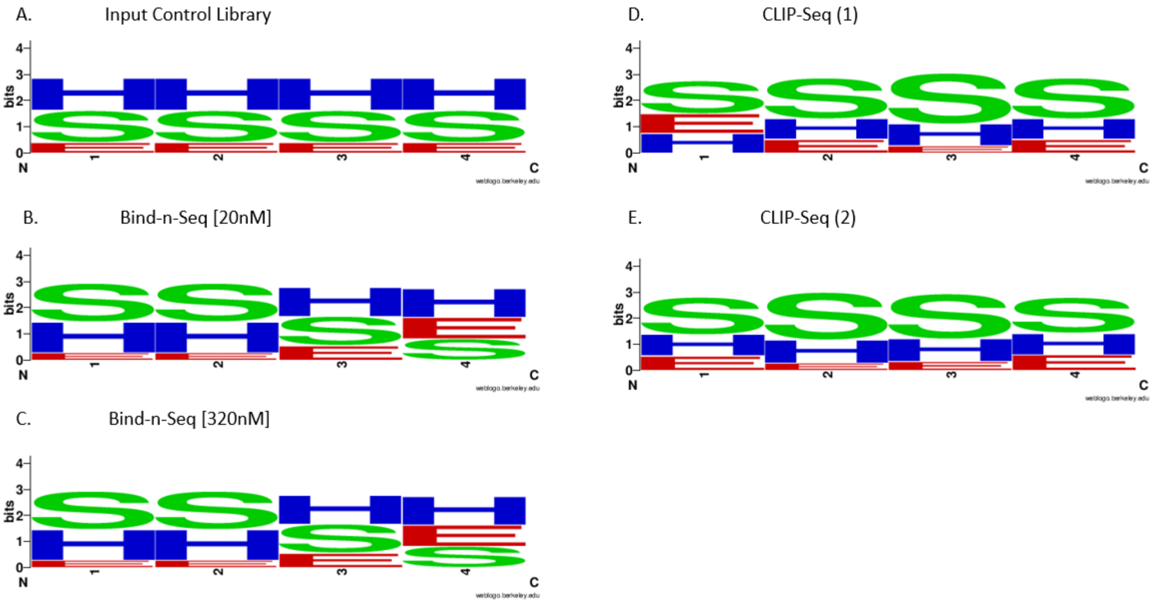


Figure 4: Secondary motifs constructed from $n = 10000$ sampled for (A) Bind-n-Seq input library control, (B) Bind-n-Seq 20nM RBP concentration, (C) Bind-n-Seq 320nM RBP concentration, (D) CLIP-Seq (first dataset), (E) CLIP-Seq (second dataset).

In both the input library and the CLIP-Seq data, equal probabilities for each of the structural elements across all positions of the motif are observed. It was also observed that the most likely elements at the first position differed between the two CLIP-Seq datasets. The two concentrations of protein in the Bind-n-Seq datasets show a more nuanced structural motif, with extreme consistency between the two datasets. The motifs 'SSHH' and 'HHSS/HHSE' are both consistent with the beginning and end of single hairpin loop structures.

Primary Sequence Motif Generation Primary sequence motifs were generated simultaneously to the secondary structure motifs, and passed into WebLogo.

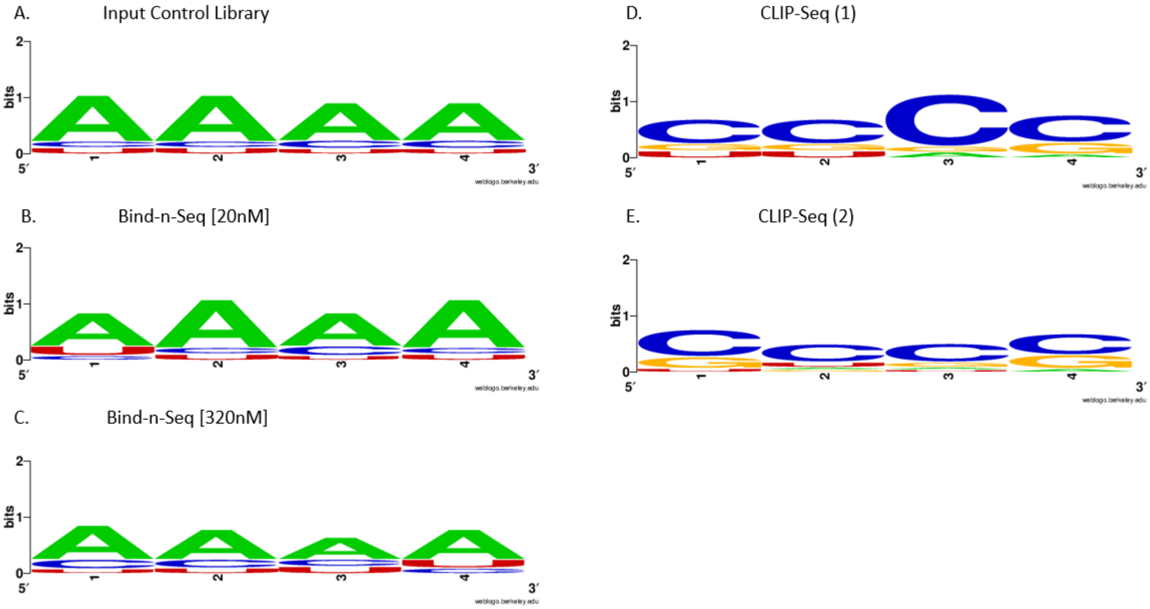


Figure 5: Primary sequence motifs constructed from $n = 10000$ sampled for (A) Bind-n-Seq input library control, (B) Bind-n-Seq 20nM RBP concentration, (C) Bind-n-Seq 320nM RBP concentration, (D) CLIP-Seq (first dataset), (E) CLIP-Seq (second dataset).

The most common base pairs appeared to differ between the bind-n-seq library and experimental data and the CLIP-Seq data. This is likely artifact from the cell samples versus the input library chosen to perform in the bind-n-seq experiments, and not necessarily indicative of any sequence motifs derived from the data.

Discussion

Due to constraints on computational power available, a motif size of 4 was used for both sequence motifs and structural motifs, compared to 6 for primary sequence and 8 for secondary structure used in Zhang et al. Additionally, only 100,000 out of 20-30 million reads from each dataset could be used due to memory constraints. It was not considered likely that any meaningful data would be obtained given these constraints, but it seems that even with this very parsimonious data that some information about structural binding preferences of IGF2BP2 could be obtained from the Bind-n-Seq data, but not from the CLIP-Seq, whose structural profile resembled the control. The Bind-n-Seq technique was developed to remove the significant noise present in CLIP-Seq experiments, and the data obtained from this project seem to support this idea. The reconstruction error of secondary structural profiles was significantly lower in Bind-n-Seq compared to CLIP and the control, even at higher protein concentrations (320 nM), which may have increased non-specific binding. Additionally, even with a small motif size of 4, some structural motif discovery appeared to be possible in the Bind-n-Seq data that was not possible in CLIP.

Conclusions Deep learning techniques are generally very computationally intensive, but it seems possible that using a model similar to the one implemented in this project, larger structural motifs could be determined. There have yet to be significant deep learning attempts at RBP motif discovery using Bind-n-Seq data, and the reduced noise from these datasets may improve model accuracy, and provide more information about transcriptionally relevant motifs.

References

[1] M. A. Reynoso, P. Juntawong, M. Lancia, F. A. Blanco, J. Bailey-Serres, and M. E. Zanetti, "Translating ribosome affinity purification (trap) followed by rna sequencing technology (trap-seq) for quantitative assessment of plant translatomes," *Methods Mol Biol*, vol. 1284, pp. 185–207, 2015. Reynoso, Mauricio A Juntawong, Piyada Lancia, Marcos Blanco, Flavio A Bailey-Serres, Julia Zanetti, Maria Eugenia eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2015/03/12 06:00 Methods Mol Biol. 2015;1284:185-207. doi: 10.1007/978-1-4939-2444-89.

[2] N. Lambert, A. Robertson, M. Jangi, S. McGeary, P. A. Sharp, and C. B. Burge, "Rna bind-n-seq: quantitative assessment of the sequence and structural binding specificity of rna binding proteins," *Mol Cell*, vol. 54, no. 5, pp. 887–900, 2014. Lambert, Nicole Robertson, Alex Jangi, Mohini McGeary, Sean Sharp, Phillip A Burge, Christopher B eng T32 GM007287/GM/NIGMS NIH HHS/ P01 CA042063/CA/NCI NIH HHS/ P30 CA014051/CA/NCI NIH HHS/ U54 HG007005/HG/NHGRI NIH HHS/ T32 GM087237/GM/NIGMS NIH HHS/ R01 GM034277/GM/NIGMS NIH HHS/ R01 HG002439/HG/NHGRI NIH HHS/ R01 GM085319/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. 2014/05/20 06:00 Mol Cell. 2014 Jun 5;54(5):887-900. doi: 10.1016/j.molcel.2014.04.016. Epub 2014 May 15.

[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 Res*, vol. 34, no. 17, p. e117, 2006. Hiller, Michael Pudimat, Rainer Busch, Anke Backofen, Rolf eng Evaluation Studies Research Support, Non-U.S. Gov't England 2006/09/22 09:00 Nucleic Acids Res. 2006;34(17):e117. doi: 10.1093/nar/gkl544. Epub 2006 Sep 20.

[4] X. Pan, P. Rijnbeek, J. Yan, and H. B. Shen, "Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks," *BMC Genomics*, vol. 19, no. 1, p. 511, 2018. Pan, Xiaoyong Rijnbeek, Peter Yan, Junchi Shen, Hong-Bin eng 61671288/National Natural Science Foundation of

China 61462018/National Natural Science Foundation of China England 2018/07/05 06:00 BMC Genomics. 2018 Jul 3;19(1):511. doi: 10.1186/s12864-018-4889-1.

[5] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure," *Proc Natl Acad Sci U S A*, vol. 101, no. 19, pp. 7287–92, 2004. Mathews, David H Disney, Matthew D Childs, Jessica L Schroeder, Susan J Zuker, Michael Turner, Douglas H eng T32 GM007356/GM/NIGMS NIH HHS/ GM 22939/GM/NIGMS NIH HHS/ 5T32 GM 07356/GM/NIGMS NIH HHS/ R01 GM022939/GM/NIGMS NIH HHS/ GM 54250/GM/NIGMS NIH HHS/ T32 DE 07202/DE/NIDCR NIH HHS/ T32 DE007202/DE/NIDCR NIH HHS/ R01 GM054250/GM/NIGMS NIH HHS/ Research Support, U.S. Gov't, P.H.S. 2004/05/05 05:00 Proc Natl Acad Sci U S A. 2004 May 11;101(19):7287-92. doi: 10.1073/pnas.0401799101. Epub 2004 May 3.

[6] I. Ben-Bassat, B. Chor, and Y. Orenstein, "A deep neural network approach for learning intrinsic protein-rna binding preferences," *Bioinformatics*, vol. 34, no. 17, pp. i638–i646, 2018. Ben-Bassat, Ilan Chor, Benny Orenstein, Yaron eng England 2018/11/14 06:00 Bioinformatics. 2018 Sep 1;34(17):i638-i646. doi: 10.1093/bioinformatics/bty600.

[7] S. Zhang, J. Zhou, H. Hu, H. Gong, L. Chen, C. Cheng, and J. Zeng, "A deep learning framework for modeling structural features of rna-binding protein targets," *Nucleic Acids Res*, vol. 44, no. 4, p. e32, 2016. Zhang, Sai Zhou, Jingtian Hu, Hailin Gong, Haipeng Chen, Ligong Cheng, Chao Zeng, Jianyang eng P30 CA023108/CA/NCI NIH HHS/ Research Support, Non-U.S. Gov't England 2015/10/16 06:00 Nucleic Acids Res. 2016 Feb 29;44(4):e32. doi: 10.1093/nar/gkv1025. Epub 2015 Oct 13.

[8] F. A. I. C., "An introduction to restricted boltzmann machines," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, vol. 7441 of Lecture Notes in Computer Science, 2012. Berlin; Heidelberg Springer 14 36.

[9] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Viennarna package 2.0," *Algorithms Mol Biol*, vol. 6, p. 26, 2011. Lorenz, Ronny Bernhart, Stephan H Honer Zu Siederdissen, Christian Tafer, Hakim Flamm, Christoph Stadler, Peter F Hofacker, Ivo L eng England 2011/11/26 06:00 Algorithms Mol Biol. 2011 Nov 24;6:26. doi: 10.1186/1748-7188-6-26.

[10] H. G.E., "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, pp. 1771–1800, 2002.