Correlated evolution-guided binding sequence specificity predictions for experimentally unexplored RNA-binding proteins

Shu Yang[1]*✉, Jiahang Sha[1]*, Kefei Liu[1], Sumita Garai[1], Jingxuan Bao[1], Zixuan Wen[1], Raymond T. Ng[2] and Li Shen[1]✉

[1] Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

[2]Department of Computer Science, University of British Columbia, Vancouver, Canada

*These authors contributed equally to this work

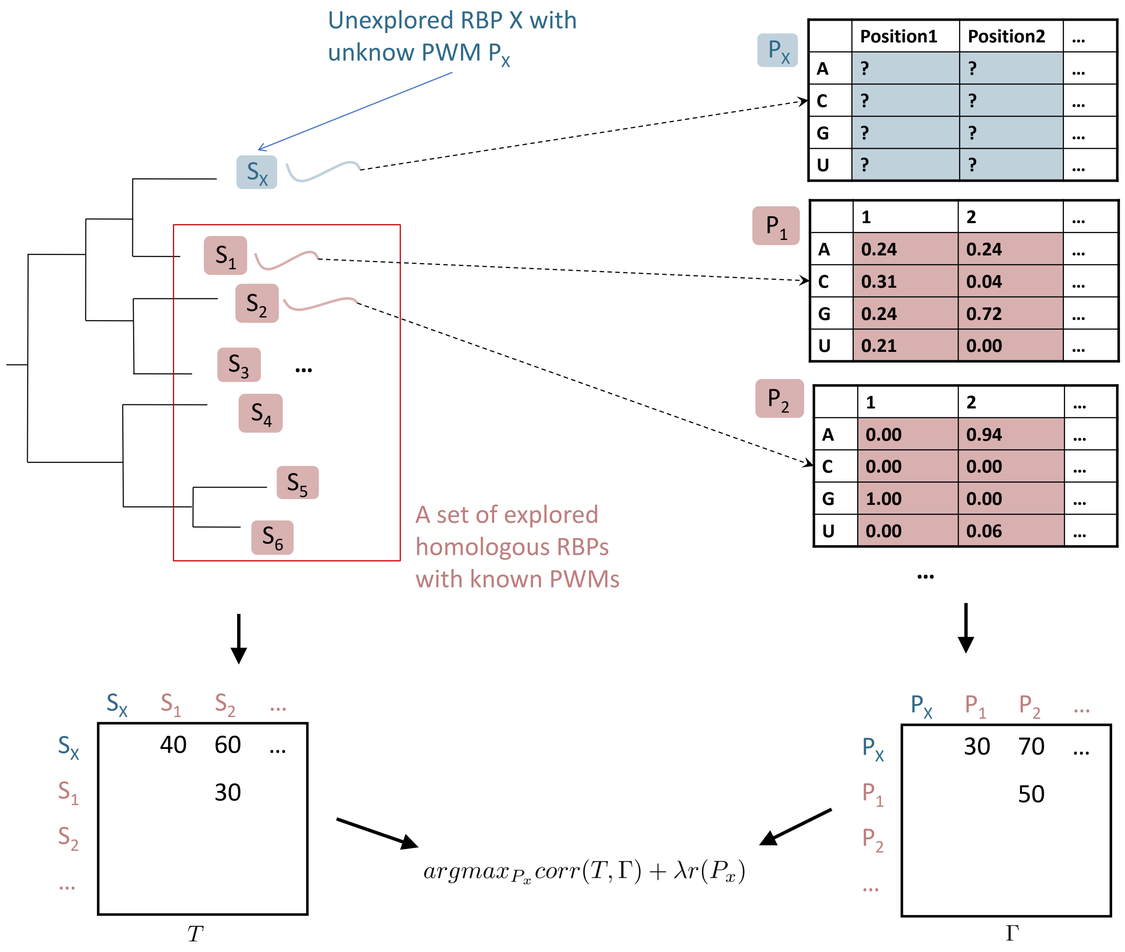
✉To whom correspondence should be addressed: Li.Shen@pennmedicine.upenn.edu, Shu.Yang@pennmedicine.upenn.edu

Introduction

- Characterizing the binding preferences of RNA binding proteins (RBPs) is essential to decipher the underlying mechanism of the cross-talk between RNAs and proteins.
- Many RBPs are reported to display intrinsic preferences to specific sequence motifs, although RNAs can fold into various structures in the cell.
- Existing methods were mostly focusing on deriving the binding specificity of particular proteins from their experimental data like CLIP, or RNAcompete, etc.
- However, only a small subset of RBPs have their binding specificities experimentally explored; the binding information on the vast majority of the RBPs is still unknown.
- Here, we propose to utilize the correlated evolutionary relationship between RBPs and their target RNA sequence motifs to formulate the prediction of sequence specificity as an optimization to maximize the correlated evolution.

Methods

We took Position Weight Matrix (PWM) as our sequence specificity representation and predicted the PWM for the unexplored RBP based on correlated evolution.



Algorithm 1: PWM prediction based on correlated evolution

Data: PWMs: $\{\mathbf{P_1} \in \mathbb{R}^{4\times L}, ..., \mathbf{P_{n-1}} \in \mathbb{R}^{4\times L}\}$, RBP sequence similarity matrix: $\mathbf{T} \in \mathbb{R}^{n\times n}$ derived from MSA of sequences $\{\mathbf{S_X}, \mathbf{S_1}, ..., \mathbf{S_{n-1}}\}$

Result: Predicted $\mathbf{P_X}$ for RBP X

Initialize $\mathbf{P_X}$, λ ;

$\mathbf{P_X} \leftarrow$ PROJ_TO_SIMPLEX($\mathbf{P_X}$) ; `/* probability simplex */`

repeat

  Compute the similarity matrix $\mathbf{\Gamma} \in \mathbb{R}^{n\times n}$ for $\{\mathbf{P_X}, \mathbf{P_1}, ..., \mathbf{P_{n-1}}\}$;

  Compute $d(\mathbf{T}, \mathbf{\Gamma}) + \lambda r(\mathbf{P_X})$ with a metric d and regularization r of choice;

  Descend $d(\mathbf{T}, \mathbf{\Gamma}) + \lambda r(\mathbf{P_X})$ with respect to $\mathbf{P_X}$;

  $\mathbf{P_X} \leftarrow$ PROJ_TO_SIMPLEX($\mathbf{P_X}$);

until *convergence*;

return $\mathbf{P_X}$

Dataset

The primary dataset containing protein sequences and PWMs was collected from cisBP-RNA database (http://cisbp-rna.ccbr.utoronto.ca/) which is based on binding data from RNAcompete in vitro assay. Here, we used a subset of RBPs from the RRM (RNA Recognition Motif) family.

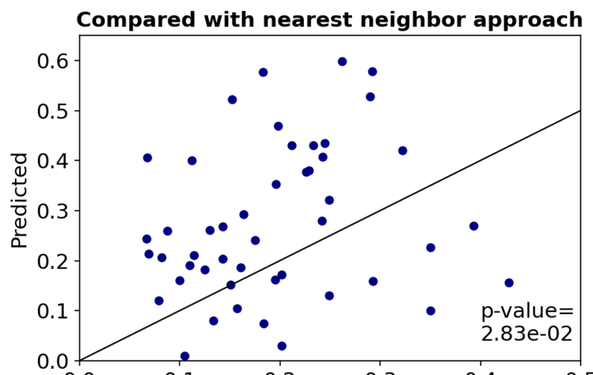
RBP family	# of proteins	RNAcompete[1]	Correlation
RRM	51	Full length	0.419**

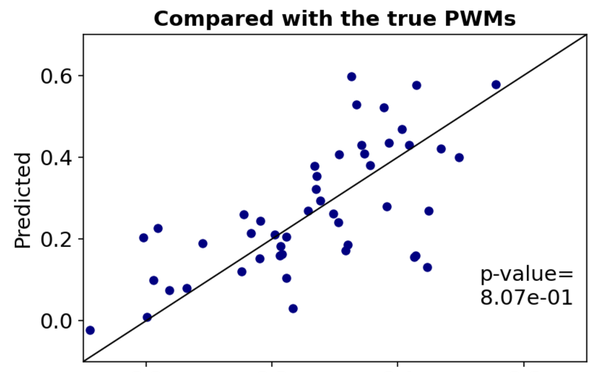
[1]: RNAcompete protein construct

**: p-value < 0.01 based on a two-tailed test against the null hypothesis that correlation = 0

Results

- Following previous studies to benchmark the performance: our input PWMs were derived by cisBP-RNA on their RNAcompete probes setA, and our predicted PWMs were evaluated on another separate RNAcompete probes setB.
- We evaluated on the RRM dataset in a leave-one-out manner, and we compared with the nearest neighbor approach (left) cisBP-RNA uses when infer PWMs for unexplored RBPs as well as the true PWMs (right) of the predicted RBPs .





- The results showed that our predictions were significantly better than the previous nearest neighbor alternative (left) and comparable to the true specificity models (right).

Discussion and Conclusion

- As the experimental characterization of the binding specificities of RBPs is not always available, our method may serve as a simple workaround to provide insights on the bindings of unexplored RBPs.
- For the next step, we plan to extend our analysis to other RBP families in the cisBP-RNA database and also apply the method to CLIP datasets which capture the in vivo binding.

Acknowledgements:

This work was supported in part by the National Institutes of Health [R01 AG071470, U01 AG068057, R01 AG066833].