Predicting RNA Protein Binding Sites

Eman Ansar and Sara Zewil,
Third year, Computer Science,
Carnegie Mellon University-Qatar

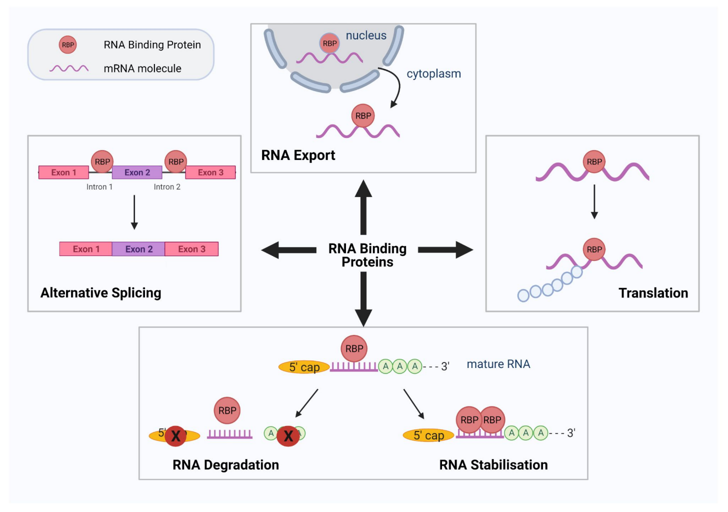
Presentation Outline

01	02	03
Why Predict RNA Protein Binding Sites ?	How to Capture these Complex Patterns	State of the Art
04	**05 & 06**	**07**
Methodology	Results & Conclusion	Teamwork & Acknowledgement

Why Predict RNA Protein Binding Sites ?

RNA-binding proteins are key players in cellular processes

- Importance
  - RNA splicing, localization and stabilization
- Challenges
  - Complex structure

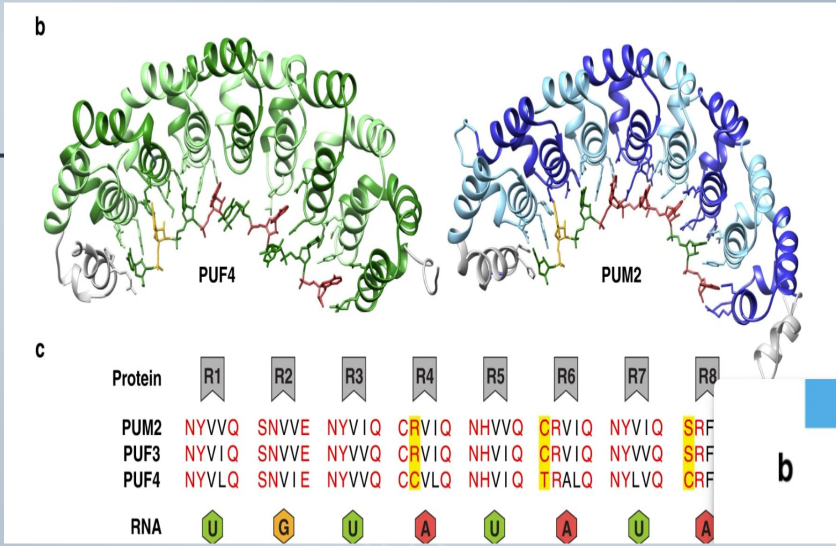


RBPs take over 5–10% of the eukaryotic proteome and play key roles in many biological processes, yet not accurately predicted

How to Capture these Complex Patterns

Hard because of complex structures and preferences

- For a specific RNA-binding protein
  - Given RNA sequence + binding preferences
- Predict the potential binding sites within RNA.



Protein	R1	R2	R3	R4	R5	R6	R7	R8
PUM2	NYVVQ	SNVVE	NYVIQ	CRVIQ	NHVVQ	CRVIQ	NYVIQ	SRF
PUF3	NYVIQ	SNVVE	NYVVQ	CRVIQ	NHVIQ	CRVIQ	NYVVQ	SRF
PUF4	NYVLQ	SNVIE	NYVVQ	CCVLQ	NHVIQ	TRALQ	NYLVQ	CRF
RNA	U	G	U	A	U	A	U	A

Current methods are both time-consuming and costly due to the need for extensive experimental validation and computational analysis.

State of the Art

RNAContext

Identifies RBP-specific preferences in both sequence and structure
Kazan et al., 2010

GraphProt

Uses graph encoding of RNA sequences and SVM classification for RNA binding prediction
Maticzka et al., 2014

RNAcommender

Built on a recommender system to predict RBP binding targets by using protein domain composition and predicting the RNA’s secondary structures
Corrado et al., 2016

iONMF

Improves prediction accuracy by integrating various data sources like sequences, structures, gene types, and CLIP co-binding through orthogonal matrix factorization.

DeepBind/DeepSea

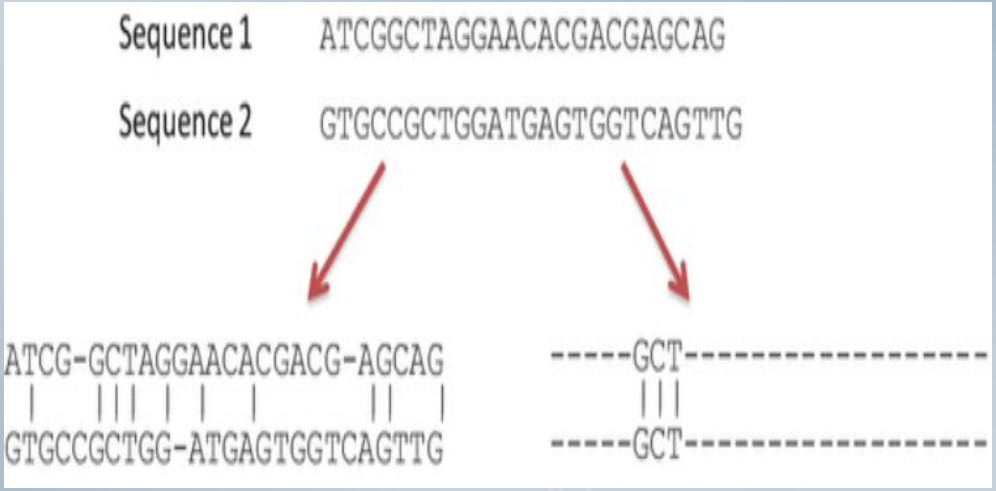
Demonstrate superior prediction accuracy compared to older techniques, capturing complex patterns in the data
Alipanahi et al., 2015; Zhou & Troyanskaya, 2015

CNNs

They possess the unique capability to automatically extract valuable binding motifs crucial for understanding RBP-RNA interactions
LeCun et al., 1998

Local & Global RNA Sequences

- Why consider local ?
  - the specificity of protein-RNA interactions.
  - selective affinity an RBP has for its RNA target.



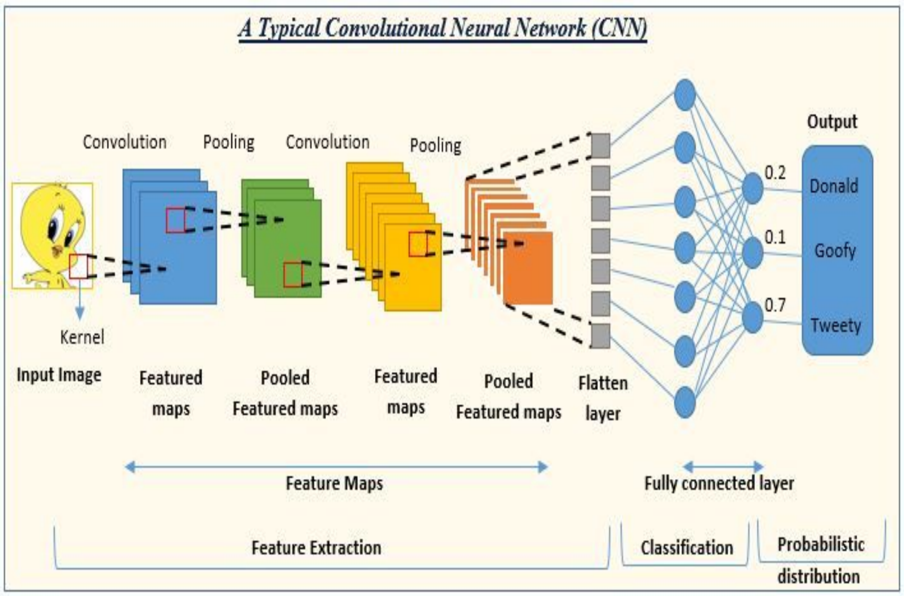
```
Sequence 1    ATCGGCTAGGAACACGACGAGCAG

Sequence 2    GTGCCGCTGGATGAGTGGTCAGTTG
```

```
ATCG-GCTAGGAACACGACG-AGCAG        -----GCT------------------
 |   ||| | |  |      ||  |             |||
GTGCCGCTGG-ATGAGTGGTCAGTTG        -----GCT------------------
```

Convolutional Neural Networks

Hard because of complex structures and preferences

- Specialized for processing structured grid data
  - Convolutional, pooling, and fully connected layers.
- Robust to variations in scale, orientation, and lighting.sites.



They are ideal for RNA sequences due to their capability to automatically learn and capture spatial patterns, including local and global structural features.

Methodology

Dataset

RBP-24 Dataset: It centers on the intricate details of how proteins interact with RNA at a subsequence level.

Processing Global

For Global RNA sequences we padded all sequences to have the same length of the longest RNA sequence.

Processing Local

We had to divide the sequences into multiple subsequences using window size W, each subsequence is a channel , and then decide the maximum number of channels using the maximum length sequence in the training sequences.

Training

- Train separate 24 Protein Files and predict separately for each.

Parameter Tuning

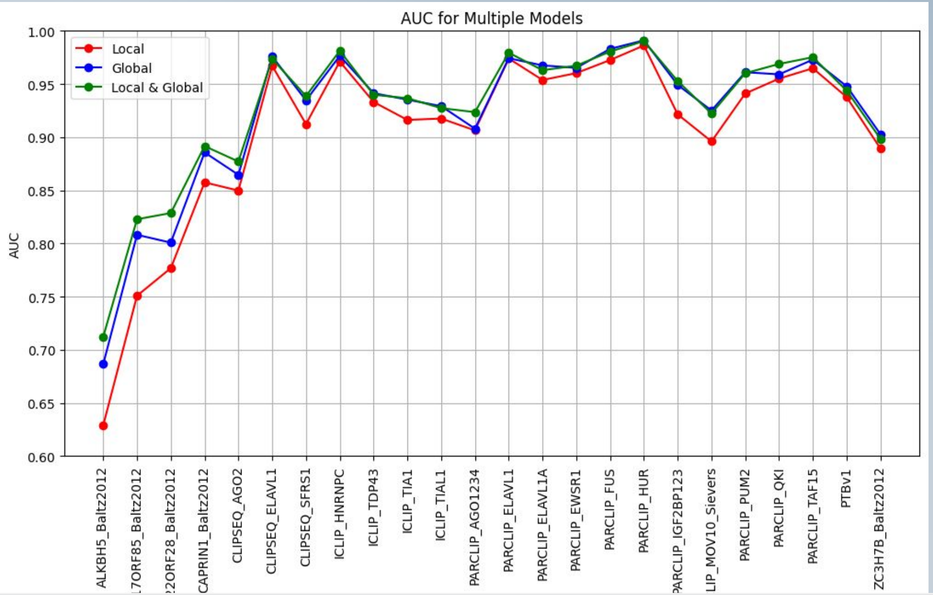
Used the ones in the paper determined by grid search. Window size W = 101, kernel size for the first convolution layer = (4, 10), kernel size for the second convolution layer = (1,10), stride S = (1,1) and pool size = (1,3).

Getting Results

Using matplotlib and AUC calculation.

Results

- Parameters Optimization
- Performance Metric
  - average AUC scores
- Performance on RBP-24
  - 0.906 for Local
  - 0.922 for Global
  - 0.927 for Integration



All in all, we were able to replicate the results and show that incorporating both gives us better accuracy in predictions.

Conclusion

- Enhancing Predictive Models
  - Considering global and local sequences
- Improvements
  - More datasets for validation

All in all, we were able to replicate the results and show that incorporating both gives us better accuracy in predictions.

Team Work & Acknowledgement

- EA: Software, Writing-review & editing
- SZ:Writing-original draft & Visualization

Thanks for the shared data from GraphProt, and Xiaoyong Pan for providing the data.

(email: 2008xypan@sjtu.edu.cn)

References

1.Maticzka, D. et al. (2014) ‘GraphProt: Modeling binding preferences of RNA-binding proteins’, Genome Biology, 15(1). doi:10.1186/gb-2014-15-1-r17. Author 1, A.; Author 2, B. Title of the chapter. In Book Title, 2nd ed.; Editor 1, A., Editor 2, B., Eds.; Publisher: Publisher Location, Country, 2007; Volume 3, pp. 154–196.
2.Alipanahi, B. et al. (2015) ‘Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning’, Nature Biotechnology, 33(8), pp. 831–838. doi:10.1038/nbt.3300. Author 1, A.B.; Author 2, C. Title of Unpublished Work. Abbreviated Journal Name year, phrase indicating stage of publication (submitted; accepted; in press).
3.Pan, X. and Shen, H.-B. (2018) ‘Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional Neural Networks’, Bioinformatics, 34(20), pp. 3427–3436. doi:10.1093/bioinformatics/bty364.
4.GraphProt - modeling binding preferences of RNA-binding proteins (no date) Bioinformatics Group Freiburg - GraphProt - modeling binding preferences of RNA-binding proteins. Available at: http://www.bioinf.uni-freiburg.de/Software/GraphProt
5.Quang, D. and Xie, X. (2016) ‘DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences’, Nucleic Acids Research, 44(11). doi:10.1093/nar/gkw226.
6.Zhou, J. and Troyanskaya, O.G. (2015) ‘Predicting effects of noncoding variants with deep learning–based sequence model’, Nature Methods, 12(10), pp. 931–934. doi:10.1038/nmeth.3547. Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.