genes

MDPI

Method Paper

Predicting RNA - protein binding sites using local and global deep Convolutional Neural Networks

Eman Ansar [1], and Sara Zewil [2]

[1] CMU-Q; eansar@andrew.cmu.edu
[2] CMU-Q; szewil@andrew.cmu.edu

Abstract: RNA-binding proteins (RBPs) play a crucial role in a range of cellular processes, including RNA splicing, stabilization, and localization, by binding to specific RNA sequences or structures. The identification of RNA-protein binding sites is key to understanding gene expression regulation and has traditionally relied on labor-intensive experimental methods. Given the scalability challenges of such methods, computational models, notably deep learning techniques, have emerged as promising alternatives. However, so far these techniques have only used global sequences of RNAs. Local sequences are also known to have a great biological significance in the RNA-protein interactions, as they are integral in determining the specificity of binding sites, thus enriching our understanding of protein-RNA affinities. Our work propels the integration of both local and global convolutional neural networks (CNNs) to discern structural attributes of RNA, facilitating a comprehensive analysis of RNA-protein interactions. Utilizing the CLIP-seq (Cross-linking and Immunoprecipitation) dataset RBP-24, which focuses on subsequence-level RNA-protein interactions, wherein proteins selectively interact with distinct regions or sequences embedded within extensive RNA molecules. This approach aims to enhance the accuracy of RNA binding sites predictions by integrating local and global RNA sequences. The potential outcomes of this research span from novel insights into disease mechanisms and drug discovery to a better understanding of gene regulation, ultimately contributing to the advancement of molecular biology and its applications in health and disease treatment. From our results, we see that our approach does enhance predictive when compared with using just the global or local CNNs on their own.

Keywords: Convolutional Neural Networks; RNA protein binding sites; nucleotides; motifs; CLIP-seq; RBP-24; specificities; local sequences; global sequences; AUC

1. Introduction

RNA-binding proteins (RBPs) are key players in a wide array of cellular processes, including RNA splicing, stabilization, and the localization of RNA within cells. These proteins bind to specific RNA sequences or motifs, playing a critical role in the regulation of gene expression. Given their importance, the accurate prediction of RNA-protein binding sites and motifs is a critical scientific challenge. Identifying these sites involves understanding a complex mix of sequence patterns, RNA shapes, and biochemical properties that make every binding site unique, alongside considering the shape and function of the proteins. This specificity means that each binding site is suited for a particular protein.

One approach to identifying RNA-protein binding sites has been through labor-intensive and costly experimental methods, which are impractical for large-scale analysis. Given these challenges, computational models, notably deep learning techniques, have emerged as promising alternatives.

The contemporary state-of-the-art in the domain of RNA-binding site prediction is characterized by an eclectic application of machine learning paradigms, encompassing methodologies such as Support Vector Machines (SVMs), Random Forests, and advanced Deep Learning algorithms. Concurrent with these computational techniques are strategies derived from structural bioinformatics, notably molecular docking, which provide alternative vantage points for the examination of RNA-protein interactions. Among these methodologies, machine learning frameworks like GraphProt, which leverage SVMs (Maticzka et al., 2014), have significantly contributed to the progression of this research area. Nonetheless, it is the integration of Deep Learning approaches, particularly those harnessing Convolutional Neural Networks (CNNs), that have markedly enhanced the predictive accuracy and motif detection efficacy within this field. Exemplary implementations such as DeepBind (Alipanahi et al., 2015) epitomize the profound impact of CNNs, heralding a new era in the precise delineation and comprehensive understanding of RNA-protein binding dynamics.

However, so far these techniques have only used global sequences of RNAs and overlooked the role of local sequences—specific segments within the RNA that may hold the key to understanding the specificity of protein-RNA interactions. These local sequences often contain motifs or structural elements that are critical for the recognition process, shaping the selective affinity an RBP has for its RNA target.

We are reimplementing Pan and Shen (2018) CNNs that focuses on improving these predictions by using convolutional neural networks (CNNs), that consider both the global and the local sequences of RNA through integrating both global and local CNNs.

This approach aims to enhance the accuracy of RNA binding sites predictions by integrating local and global RNA CNNs. The potential benefits of this research are vast, ranging from better insights into disease mechanisms and aiding in drug discovery to improving our understanding of how genes are regulated. By offering a more comprehensive way to predict RNA-protein binding sites, we hope to contribute to the broader field of molecular biology, enabling further discoveries and applications that could have a direct impact on health and disease treatment strategies.

2. Materials and Methods

2.1 Data

This work utilizes CLIP-seq (Cross-linking and Immunoprecipitation) dataset to study protein interactions with RNA.

RBP-24 Dataset: It centers on the intricate details of how proteins interact with RNA at a subsequence level. This means that the dataset provides granular insights into the ways in which proteins recognize and bind to specific regions or sequences within the larger RNA molecules. This selective binding is critical for understanding the specificity of protein-RNA interactions, as it influences various cellular processes by modulating the function of the bound RNA molecules. The training and testing data for RBP-24 binding sites are sourced from the GraphProt website (http://www.bioinf.uni-freiburg.de/Software/GraphProt). This dataset includes 24 experiments featuring 21 RBPs. In each experiment, positive sites refer to specific subsequences within RNA molecules. These subsequences are identified as RNA binding sites and are determined based on peak centers derived from CLIP-seq data. The peak center is a characteristic point within a peak region indicating the most probable binding site. Conversely, negative sites denote regions within RNA molecules where there is no supportive evidence of them being RNA binding sites.

2.2 Method

2.2.1. Preprocessing

To begin with we started by data preprocessing, where CNNs input should be sequences of fixed length and since RNA sequences might vary in length, we padded the sequences so that they all have the same length. We used the nucleotide ‘N’ for padding.

For Global RNA sequences we padded all sequences to have the same length of the longest RNA sequence, While for local sequences we had to divide the sequences into multiple subsequences using window sixe W, each subsequence is a channel , and then decide the maximum number of channels using the maximium length sequence in the training sequences.

For a local RNA sequence of length L, is subdivided into multiple subsequences. Each of these subsequences is then treated as an independent channel. This segmentation process utilizes a window size parameter denoted as W, allowing for a systematic partitioning of the sequence. The determination of the total number of subsequences across the entire sequence accounts for any overlapped shift, denoted by S, and can be calculated as (L—W)/(W—S). Moreover, to ensure consistency and compatibility within the local CNN framework, it is imperative to establish a maximum number of channels, denoted as C. This determination is based on an analysis of the longest sequence observed within the training dataset. In instances where the number of channels for a specific sequence is less than C, a supplementation strategy is implemented. This involves the addition of supplementary channels derived from sequences containing exclusively nucleotide Ns. The augmentation process continues until the maximum channel count, C, is reached.

After padding, the nucleotides are encoded to one-hot encoding matrix, as described by Alipanahi et al. (2015), Quang and Xie (2016), and Zhou and Troyanskaya (2015). involves transforming an RNA sequence $(s_1, s_2, \dots, s_n)$ with n nucleotides, and a motif detector with a convolutional filter kernel size m of convolve filters. This matrix, denoted as M, is defined as follows:

$$Mi,j = \begin{cases} 0.25 & if\ si - m + 1 = N\ or\ i < m\ or\ i > n - m \\ 1 & if\ si - m + 1\ in (A, C, G, U) \\ 0 & otherwise \end{cases}$$

2.2.2. Convolutional neural network and long short-term memory network

The Convolutional neural network (CNN) architecture is composed of 2 convolution layers, max pooling, and fully connected layers. The convolution layer applies a set of filters across the input one-hot matrix to compute the pointwise product between the input and each filter, yielding a series of convolutional feature maps. These maps are then processed through a Rectified Linear Unit (ReLU), which retains only the positive values, allowing the CNN to emphasize significant features in the RNA sequences. Subsequently, max-pooling layer is applied to reduce the feature maps' dimensionality by selecting the maximal value within a specified window. We then apply the fully connected layers. However, to avoid overfitting we add dropout layers.

3. Results

We trained 3 CNNs. The first is the local CNN that uses the local sequences to predict binding sites in RNA, the second is the Global CNN where it uses the global RNA sequences for prediction while the last one integrates both the global and local sequences. For the local CNN we got a mean AUC score of 0.906, while global CNN score is 0.922 and for the combined CNN the mean AUC score was 0.927.

3.1. Parameter optimization

We followed the same parameters specified in which they found using grid search. The parameters used were as follows: 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).

3.2 The performance of local and Global CNNs on RBP-24

All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

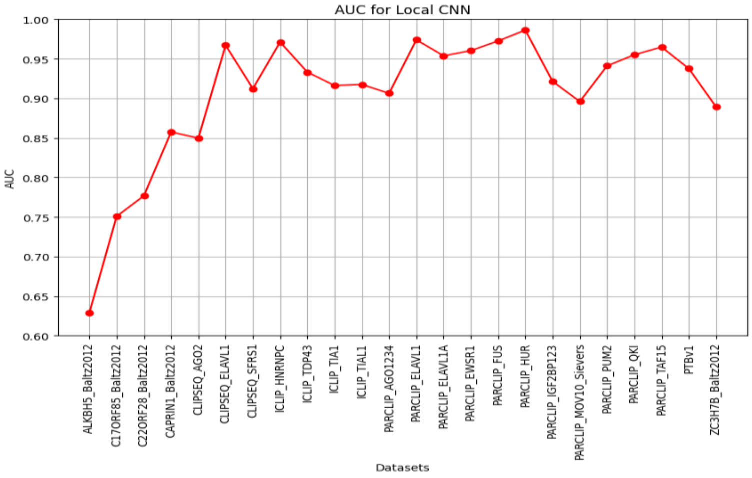
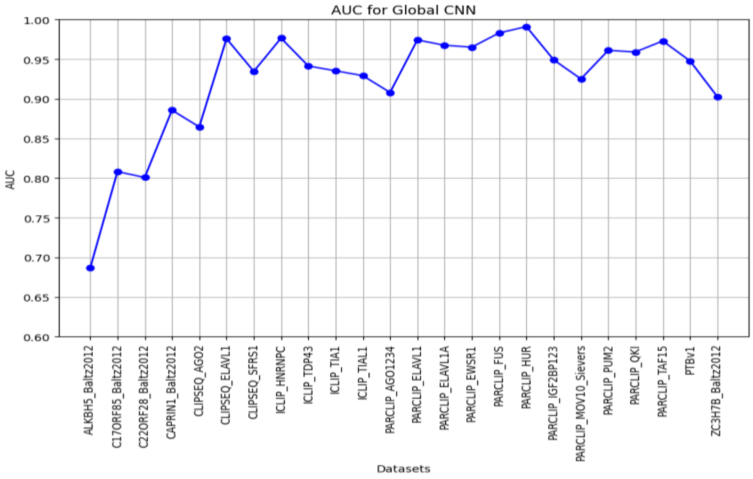


Figure 1. This figure shows AUC plots for the local CNNs.



\

Figure 2. This figure shows AUC plots for the global CNNs.

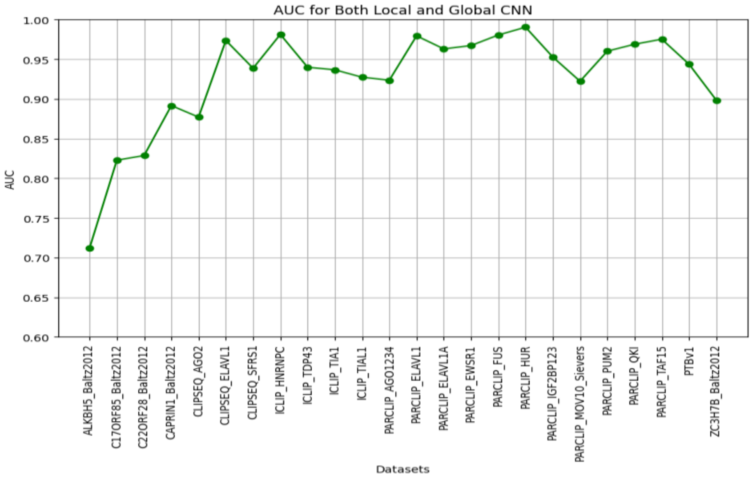


Figure 3 This figure shows AUC plots for both local & global CNNs.

4. Discussion

The results obtained from our study demonstrate the effectiveness of utilizing both local and global convolutional neural networks (CNNs) to capture RNA-protein binding sites. Our analysis revealed that employing only local CNNs yielded an area under the receiver operating characteristic curve (AUC) of 0.906, while using solely global CNNs resulted in an AUC of 0.922. However, when we combined the context from both the local and global sequences, the AUC significantly increased to 0.927.

This finding underscores the importance of considering both local sequence motifs and global sequence features when predicting RNA-protein binding sites. Local CNNs are adept at capturing short-range interactions and sequence motifs that are characteristic of binding sites, while global CNNs can capture broader contextual information and long-range dependencies within the sequences. By integrating information from both types of networks, we can leverage the complementary strengths of each approach, resulting in improved predictive performance.

In the context of previous studies, our findings align with the understanding that RNA-protein interactions are influenced by a combination of sequence motifs and structural features. This underscores the necessity of considering multiple scales of information for accurate prediction.

Beyond RNA-protein interactions, our approach may have broader implications for computational biology, particularly in predicting other molecular interactions. Understanding the principles of molecular recognition is crucial for deciphering gene expression regulation and cellular function.

5. Conclusions

All in all, our study demonstrates the efficacy of integrating local and global convolutional neural networks (CNNs) to enhance predictive models for RNA-protein interactions. We highlight the importance of considering both local sequence motifs and global sequence features, which significantly improves predictive accuracy. Moving forward, further exploration of additional data sources and experimental validation will refine our models.

Additionally, investigating dynamic interactions and diverse cellular contexts promises deeper insights into regulatory networks. Our findings contribute to advancing our understanding of molecular recognition in cellular biology and underscore the importance of integrating computational and experimental approaches for continued progress in this field, with implications for both basic research and therapeutic development.

Author Contributions: EA; Software. SA; Writing-original draft. EA; Writing- review & editing. SA; Visualization.

Acknowledgments: Thanks for the shared data from GraphProt, and Xiaoyong Pan for providing the data. (email: 2008xypan@sjtu.edu.cn)

Conflicts of Interest: The authors declare no conflicts of interest.

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.