RNA LLM and generating RNA aptamers by artificial intelligence.

Shaoxuan Tang

Abstract

Abstract. Adapters are a special class of single-stranded nucleic acid molecules that exhibit a high degree of specificity and affinity in binding to particular target molecules. This capability has led to their widespread applications across various biomedical and biotechnological domains. In recent years, with the rapid advancement of machine learning techniques, particularly the success of Transformer models in the field of natural language processing, researchers have begun exploring the application of such models in the analysis of biological sequences. The key advantage of Transformer models lies in their ability to handle long-range dependencies, which is crucial for understanding and predicting the structure and function of large biomolecules like RNA. In summary, the combination of these two not only propels the advancement of adapter technology but also paves the way for the application of machine learning in bioinformatics.
Keywords: RNA, LLM.

1 Introduction

RNA aptamers are a special class of single-stranded RNA molecules that are selected through in vitro chemical evolution processes, such as SELEX (Systematic Evolution of Ligands by Exponential Enrichment) . They can specifically and highly selectively bind to particular targets, including proteins, small molecules, or even living cells. This binding capability arises from the three-dimensional structure of RNA molecules, allowing aptamers to recognize and tightly bind to their targets. RNA possesses high specificity and affinity, enabling aptamers to selectively identify their target molecules, making them highly valuable in molecular recognition and targeted therapy. Many aptamers can be reused without com- promising their functionality. In comparison to proteins, RNA aptamers are generally more stable and easier to synthesize and modify. As probes for biological markers, aptamers can be used to detect molecular signatures of various diseases, such as cancer and heart disease. Aptamers can serve as part of drug delivery systems to enhance drug specificity and reduce side effects. Both the aptamer itself and its binding target can form the basis for drug development. With the advancement of biotechnology, research and applications of RNA aptamers are rapidly expanding, and their potential in clinical diagnostics, treatment, and biotechnology is widely recognized. Particularly, the integration of modern molecular biology techniques and nanotechnology opens up extensive prospects for the application of RNA aptamers in precision medicine and personalized therapy. The rapid development of mRNA-based COVID-19 vaccines6 has globally heightened awareness and understanding of the advantages and potential drawbacks of mRNA technology. This significant achievement has sparked considerable anticipation within the

scientific and medical fields, as experts eagerly anticipate further evidence supporting the efficacy of mRNA-based treatments in various other medical conditions.

2 How to get the RNA embedding

We used RNA-FM model to get RNA embedding ,let it be the foundation model,to make our generation more effective.This foundation model exchange RNA sequence into (L,640),After training, RNA-FM produces a L × 640 embed- ding matrix for each RNA sequence with length L .Also, the embedding from RNA-FM can be used to infer the long non- coding RNA (lncRNA) evolutionary trend, which suggests that the evolutionary information has been learned by our model implicitly. Furthermore, models using RNA-FM embeddings could improve over state-of-the-art approaches consistently on various structural-related and functional-related downstream prediction problems, including both SARS-CoV-2 study and gene expression regulation modeling .

2.1 HOW TO DESIGN RNA APTAMERS

Designing RNA aptamers using deep learning involves leveraging neural networks to model complex relationships between RNA sequences and their binding affinities to target molecules. Gather a dataset of RNA sequences along with their corresponding binding affinities or experimen- tal outcomes. This dataset should include positive exam- ples(sequences with high affinity) and negative examples (se- quences with low affinity).Using transformer to train a Aptamers generation model.RBD scores and RNA sequence as training data.And with RaptGen experiment as a con- trast,using Transformer greedy algorithm to generate RNA sequence.Assign labels to your sequences based on their binding affinitiesSplit your dataset into training and validation sets. Train the deep learning model using the training set and validate its performance on the validation set. Adjust hyperparameters such as learning rate, batch size, and the number of layers to optimize the model's performance.Define a suitable loss function for your task. Binary cross-entropy is commonly used for binary classification tasks like binding affinity prediction.

3 RNA evolutionary information

RNA evolutionary information is also explored in our studies. We apply trajectory inference (pseudotemporal ordering), which is commonly used in single-cell transcriptomics, to a subset of lncRNA with their RNA-FM embedding as input. RNAs in the subset can be classified according to different types of species. Here we obtain their evolutionary relationship from an evolutionary study of lncRNA repertoires and expression patterns. We first generate RNA-FM embeddings for the lncRNA subset. We discover that although it is hard for RNA-FM to distinguish these RNAs into different species, the embeddings are able to present a roughly accurate evolutionary trend of different species corresponding to their ground-truth timeline. The result is surprising because we do not include such evolutionary features during training and only use the pure RNA sequences. This result testified that RNA-FM deeply mined the implicit genetic message and encoded the outputs with evolutionary information.

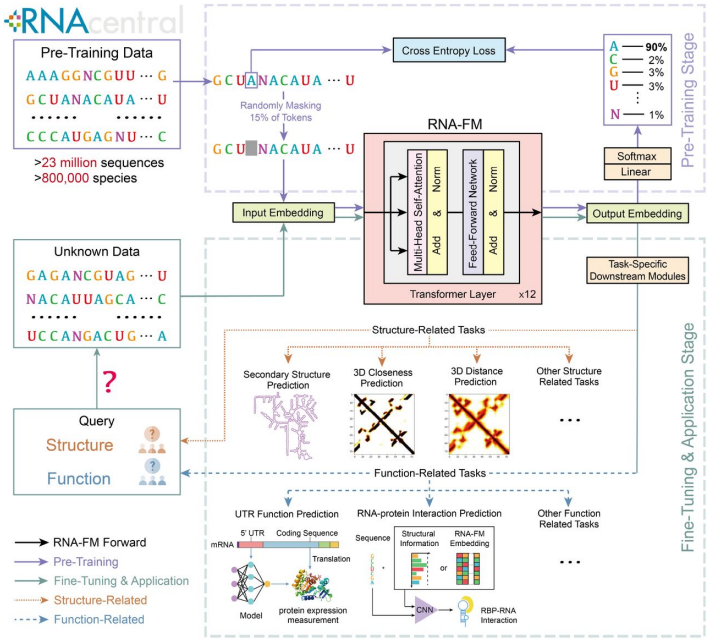

Figure 1. The RNA-FM architecture.

4 Evaluation

We look at the 10,000 RNA sequences generated to see if they are included in the SELEX sequence, and if there are none or a very small number, we increase the number generated to 100,000, and then look at the data contained in them to evaluate the accuracy of model generation and the effect of combining with RBD.Our generation strategy is to Fig. 1. The RNA-FM architecture. extract the ten sites most likely to bind to the protein from the hidden space, extract the data from them, and weight the sum to obtain the implied representation of the generated RNA aptamers.These hava several methods to evaluate RNA design. 1)Secondary Structure Prediction :Use RNA folding prediction tools to analyze the secondary structure of the designed RNA sequence. Tools like ViennaRNA, mFold, or RNAstructure can provide insights into the predicted folding patterns. 2)Functional Motifs and Targets :Check whether the designed RNA sequence contains specific functional motifs or targets necessary for its intended biological function. This may include binding sites for proteins, other RNAs, or small molecules. 3)Experimental validation is crucial to confirm the functionality of the designed RNA. This may involve in vitro assays, cell culture studies, or in vivo experiments to verify the predicted behavior

5 Conlusion

RNA design is a field of bioengineering that aims to create novel RNA molecules with desired functions in biotechnology and medicine, such as biosensing, diagnostics, and therapeu- tics. AI is starting to show its enormous potential in RNA design, such as for aptamer generation , mRNA vaccine optimization , and CRISPR/Cas-based gene editing . RNA language models can be pre-trained on a large number of RNA sequences using a variety of unsupervised learning objectives. The pre-trained RNA language model can then use supervision to fine-tune a specific task or data set. This way, the RNA language model can apply the general knowledge acquired during the pre-training stage to a specific task or dataset in the fine-tuning stage. This

enables organizations or institutions with limited computing resources to achieve effective train- ing outcomes and successful downstream task applications, thereby accelerating RNA aptamer generation and drug design. The challenge lies in identifying RNA sequences capable of folding into desired structures with functional attributes. Given the diverse application scenarios and the intricate nature of designed RNAs, conventional computational approaches for RNA structure and function design prove insufficient. Hence, there is a pressing need for further efforts to enhance the practicality and efficiency of RNA design. Employing AI- based techniques, including but not limited to deep learning, dynamic programming, and dynamic simulation, is imperative for addressing these challenges.

References

[1] Multiple sequence-alignment-based RNA language model and its application to structural inference bioRxiv 2023.03.15.532863; doi: https://doi.org/10.1101/2023.03.15.532863

[2] Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function PredictionsJiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, Yu LibioRxiv 2022.08.06.503062; doi: https://doi.org/10.1101/2022.08.06.503062

[3] UNI-RNA: UNIVERSAL PRE-TRAINED MODELS REVOLUTIONIZE RNA RESEARCHXi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, Han WenbioRxiv 2023.07.11.548588; doi: https://doi.org/10.1101/2023.07.11.548588

[4] Barbier, A.J., Jiang, A.Y., Zhang, P. et al. The clinical progress of mRNA vaccines and immunotherapies. Nat Biotechnol 40, 840–854 (2022). https://doi.org/10.1038/s41587-022-01294-2

[5] Zhang, J., Lang, M., Zhou, Y., Zhang, Y. (2023). Predicting RNA structures and functions by artificial intelligence. Trends in Genetics. https://www.cell.com/trends/genetics/fulltext/S0168-9525(23)00229-9

[6] Ryoga Ishida, Tatsuo Adachi, Aya Yokota, Hidehito Yoshihara, Kazuteru Aoki, Yoshikazu Nakamura, Michiaki Hamada, RaptRanker: in silico RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information, Nucleic Acids Research, Volume 48, Issue 14, 20 August 2020, Page e82, https://doi.org/10.1093/nar/gkaa484