**Bubble graph**

Based on the conclusion from (<https://www.cell.com/fulltext/S0092-8674(12)00536-3>), m6a sites cluster near the stop codon area most.

Therefore, we extract all known m6a sites in the 3’UTR area from our m6a sequences data and try to visualize the distribution of them. Due to the various length of the 3’UTR area of RNA sequences, we choose to use relative distance to represent the position of the sites in a uniform scale. Generally, we divide all sites into 100 bins according to their relative distance.

Then, we visualize the distribution of all m6a sites in purple, which can indicate that m6a sites cluster near the stop codon area. Then, according to the attention score, we filter out the sites visualized in green with high attention scores. Since we assume that our model can properly identify m6a sites with high attention, we calculate the proportion of sites with high attention in all m6a sites and visualize it in dark blue to eliminate the effect from the original distribution of our sample sites. According to the plot, our model properly identifies most of the known m6a sites.

Meanwhile, the radius of the bubbles represents the average attention score of the bin it is in. Our model also gives higher average attention for sites near stop codon, which may indicate that m6a sites near stop codon play more important functions than other m6a sites in biological processes.

**Tim 写的一部分：**

见学长我发给你的PDF，m6a\_experiment\_Complete\_results.pdf

**下面是之前写过的几个版本，都可以互相补充：**

Text, letter

Description automatically generated

**f1就是我们bubble图里面的紫色，f2就是淡绿色，f3就是深蓝绿色**

The aim of our study was to investigate whether our m6A modification sites prediction model, equipped with a self-attention mechanism, is capable of detecting the well-established pattern of m6A modification enrichment at the stop codon of mRNA. The self-attention mechanism in our model is designed to identify the most informative and relevant features of mRNA sequences that may harbor m6A modifications, which enhances our understanding of the post-transcriptional regulation of gene expression.

We extracted known positive samples of m6A modification sites from 3'UTR and plotted them into a density plot to analyze their distribution(f1). Since all sequences with various lengths of 3'UTR start immediately after the stop codon, we defined the relative distance of a modification site as the number of nucleotides from the stop codon to the site divided by the total number of nucleotides in the sequence. We then filtered out the sites with relatively high attention, denoted as f2. To eliminate the impact of the distribution of the original data, we divided f2 by f1 to obtain the relative percentage of sites recognized by our model in actual positive points at different relative distance, denoted as f3.

Our analysis of the density plot indicated that our model was able to identify the pattern of m6A modification enrichment at the stop codon of mRNA. Specifically, our model recognized more modification sites near the stop codon area, which is consistent with previous research demonstrating that more m6A modification sites cluster near the stop codon area. Thus, our study provides further evidence that m6A modification at the stop codon plays a critical role in post-transcriptional regulation of gene expression.

Recent research has suggested that m6A modification at the stop codon regulates the efficiency of translation termination by promoting the interaction between the mRNA and release factors. Additionally, m6A modification can affect mRNA stability by regulating the binding of RNA-binding proteins (RBPs) to the mRNA. RBPs are responsible for regulating mRNA metabolism, including stability and translation. The m6A modification at the stop codon may regulate the binding of RBPs, thus influencing mRNA stability. Furthermore, studies have shown that the m6A modification at the stop codon is involved in the regulation of alternative polyadenylation (APA), which generates multiple mRNA isoforms with different 3' UTRs by altering the site of polyadenylation. The 3' UTR plays a critical role in the regulation of gene expression, and APA can affect mRNA stability and translation efficiency. The m6A modification at the stop codon may influence APA by regulating the binding of RBPs to the mRNA.

Our investigation into the self-attention mechanism of our model provides valuable insights into the regulation of gene expression via m6A modification at the stop codon. In conclusion, our findings underscore the importance of m6A modification at the stop codon in post-transcriptional regulation of gene expression, which may lead to a better understanding of the molecular mechanisms underlying gene expression regulation.

**Filtering sequence of 3D structure**

Firstly, I extract motif in the positive sequences of our 22RBP datasets using the code of DNABERT (set the length of the motif to 7, and p value to 0.05), then I pass positive sequences of each RBP into our model and draw the landscape of each RBP. According to the high attention area visualized in the landscape, I try to find the sequence that the motif we found locates at the similar position like the landscape shown. This is because that if the motif locates at the high attention area identified by our model, it is highly possible that the motif is the sequence segment that our model recognizes as important area. In addition, since for each RBP, many motifs are extracted. What I did was choosing the most frequent to do the filtration mentioned above. (The explanation about DNABERT motif extraction, please refer the supplementary file of DNABERT)

**M6a 9 cell lines saliency and KpLogo**

Firstly, I extract motif in the positive sequences of our 9 cell lines datasets using the code of DNABERT (set the length of the motif to 7, and p value to 0.05), I try to find the sequence that the motif we found locates at the center of the sequence. This is because that data of 9 cell lines is produced by cutting the sequence with 41 window size with the m6a sites in center. In addition, since for each cell line, many motifs are extracted. What I did was choosing the most frequent to do the filtration mentioned above. (The explanation about DNABERT motif extraction, please refer the supplementary file of DNABERT). Then I feed sequences found into our model to draw saliency map (visualize attention scores).

For KpLogo, I feed the positive sequences of each cell line to KpLogo webtool to get the output.

**m6a skewing** (small graph above bubble graph)

*No window-size*

For all positive m6a sequences in our datasets, I pass each of them into our model and extract the attention score from the last layer of the model, then I calculate the real score of each nucleotide of the sequence and extract the attention scores of m6a sites and other normal sites regarding other normal sites as negative samples.

*20 window-size*

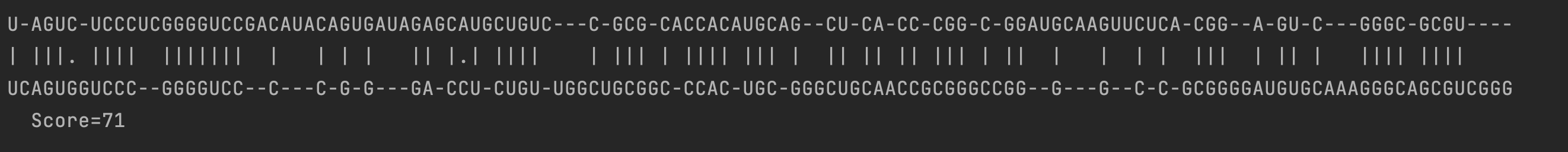
For all positive m6a sequences in our datasets, I pass each of them into our model and extract the attention score from the last layer of the model, then I use the known m6a sites as the center and crop 20 nts before and 20 nts after the center to get a segment of 41 nts in length. The attention of the m6a site is the average of attentions scores of all 41 nucleotides. This is because that we train our model using 41 nts, we use 41 as a unit to eliminate the possibility that the score of the actual sites is low, but the environment around it has high attention due to some kind of grammar existing in the environment. In addition, we choose only the non-overlap segments in our experiment. (e.g. if two 41 segments overlap, we do not use them as the data sample) For negative sites, we use the same algorithm.

Then I draw the attention score skewing density plot to show the difference of attention scores of positive and negative samples.

**Dodrio similar sequence filtering**

Firstly, I extract motif in the positive sequences of our 22RBP datasets using the code of DNABERT (set the length of the motif to 7, and p value to 0.05), then I pass positive sequences of each RBP into our model and draw the landscape of each RBP. According to the high attention area visualized in the landscape, I try to find the sequence that the motif we found locates at the similar position like the landscape shown. This is because that if the motif locates at the high attention area identified by our model, it is highly possible that the motif is the sequence segment that our model recognizes as important area. In addition, since for each RBP, many motifs are extracted. What I did was choosing the most frequent to do the filtration mentioned above. (The explanation about DNABERT motif extraction, please refer the supplementary file of DNABERT) （前置工作和前面Filtering sequence of 3D structure一样，直接复制过来了）

After this, I compare the motifs extracted by our model and the motifs in the TomTom database, and find that RBM15 is the RBP that the motif found by us match the RBM motif in the database most (对比数据库这块好像没什么意义，只是当时用错了数据库并且数据库本身的数据bias很大). Therefore, I compare the similarity of sequences filtered by the principle mentioned above in red in two ways. First, I use Biopython to align two sequences and got two sequences with relatively high alignment score in the whole sequence scale (global). Then I use BLAST (Basic Local Alignment Search Tool) to compare the similarity of sequences and found that these two sequences have a segment that is the same (local). Therefore, I choose these two sequences to draw Dodrio.

****

**Graphical user interface, application

Description automatically generatedGraphical user interface, application, Word

Description automatically generated**

**Motif Stack**

Based on the motifs extracted by DNABERT codes from 22RBP datasets, I load the PWM into ATtRACT to compare the motif we extracted with known motifs in database. Then I save the motifs that are similar to motifs that we found and match each motif with related domain and their family based on the information provided by ATtRACT database. Then I use motif stack to visualize all information.

**BertViz**

Tim help me filtering out 41 nts m6a sequence that have two m6a sites, then I try each of them to draw the BertViz, then found a relatively good one.

之前写过的一个版本：

Based on the visualization of the contextual relationship of input sequences from the m6a dataset, further analysis of 3UTRBERT model could be exerted. The attention view of heads of 3UTRBERT model are displayed in Fig.x, whose leftmost plot reveals the global self-attention pattern of an example sequence that has two m6a modification sites. Individual attentions from most 3-mer tokens in the example sequence over all heads correctly converge at the region of tokens containing the m6a modification site. By filtering tokens that contribute high attention to the m6a sites with different threshold, the relationship between the modification sites and other regions of the sequence could be better analyzed. (Fig.x right) Among all heads, heads in blue and green play important roles in heads with high attention, and those heads clearly reveal the semantic relationship within the context of the sequence specifically, since blue heads and green heads broadly highlight short regions contributing to the attention of the important token AAA with the modification site. Furthermore, the blue heads and green heads also successfully relate the token to the other modification token, demonstrating contextual understanding of the input sequence.