miRNA-mRNA interactions

lzboo 2022/9/16

I Dataset

- PAR-CLIP and CLASH experiment datasets

1.1 Train

- **site-level**

  - **positive**: 33,142

  - **negative**: 32,284

|	A	B	C	D	E	F
1	MIRNA_ID	ENSEMBL_GENE	GENE_SYMBOL	MIRNA_SEQ	MRNA_SEQ	LABEL
2	hsa-let-7a-5p	ENSG00000114573	ATP6V1A	UGAGGUAGUAGGU	GCCTGCTATTGAG(	1
3	hsa-let-7a-5p	ENSG00000104497	SNX16	UGAGGUAGUAGGU	ATTTGTTTAGTTT(	1
4	hsa-let-7a-5p	ENSG00000141682	PMAIP1	UGAGGUAGUAGGU	GCACATTGTATAT(	1
5	hsa-let-7a-5p	ENSG00000174010	KLHL15	UGAGGUAGUAGGU	GAAGTTAGACACC	1
6	hsa-let-7b-5p	ENSG00000130402	ACTN4	UGAGGUAGUAGGU	GGCCCTCATCTTC(	1
7	hsa-let-7c-5p	ENSG00000130402	ACTN4	UGAGGUAGUAGGU	GGAGGCCCTCATC	1
8	hsa-let-7a-5p	ENSG00000130402	ACTN4	UGAGGUAGUAGGU	GGAGGCCCTCATC	1
9	hsa-let-7d	ENSG00000130402	ACTN4	UGAGGUAGUAGGU	GGCCCTCATCTTC(	1
10	hsa-let-7b-5p	ENSG00000105202	FBL	UGAGGUAGUAGGU	ACCCCTTCCGCTC(	1
11	hsa-let-7c-5p	ENSG00000105202	FBL	UGAGGUAGUAGGU	ACCCCTTCCGCTC(	1
12	hsa-let-7a-5p	ENSG00000105202	FBL	UGAGGUAGUAGGU	ACCCCTTCCGCTC(	1
13	hsa-let-7d	ENSG00000105202	FBL	UGAGGUAGUAGGU	ACCCCTTCCGCTC(	1
14	hsa-let-7c-5p	ENSG00000170779	CDCA4	UGAGGUAGUAGGU	CGGTGAGCCGCAG	1
15	hsa-let-7a-5p	ENSG00000170779	CDCA4	UGAGGUAGUAGGU	CGGTGAGCCGCAG	1
16	hsa-let-7b-5p	ENSG00000164011	ZNF691	UGAGGUAGUAGGU	TATGTGCCCAGTG	1

1.2 Test

- 10 test sets

- **gene-level**

  - **positive**: 151,956

  - **negative**: 548 (randomly sampled 548 positive pairs for ten times to balance the num ber of positive and negative pairs)

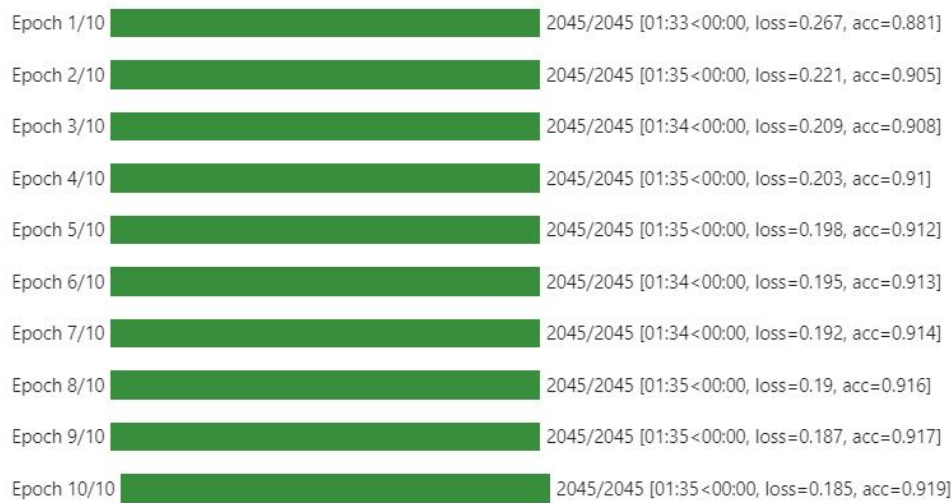|			
541	hsa-miR-99a-5p	NM_030791	0
542	hsa-miR-99a-5p	NM_031459	0
543	hsa-miR-99a-5p	NM_078467	0
544	hsa-miR-99a-5p	NM_153331	0
545	hsa-miR-99a-5p	NM_173595	0
546	hsa-miR-99a-5p	NM_181755	0
547	hsa-miR-99a-5p	NM_183078	0
548	hsa-miR-99a-5p	NM_183356	0
549	hsa-let-7a-5p	NM_032015	1
550	hsa-let-7b-3p	NM_001290257	1
551	hsa-let-7b-5p	NM_001009984	1
552	hsa-let-7b-5p	NM_001083589	1
553	hsa-let-7b-5p	NM_001145314	1
554	hsa-let-7b-5p	NM_001320923	1
555	hsa-let-7b-5p	NM_007055	1
556	hsa-let-7c-5p	NM_001130524	1

II Model

TargetFM
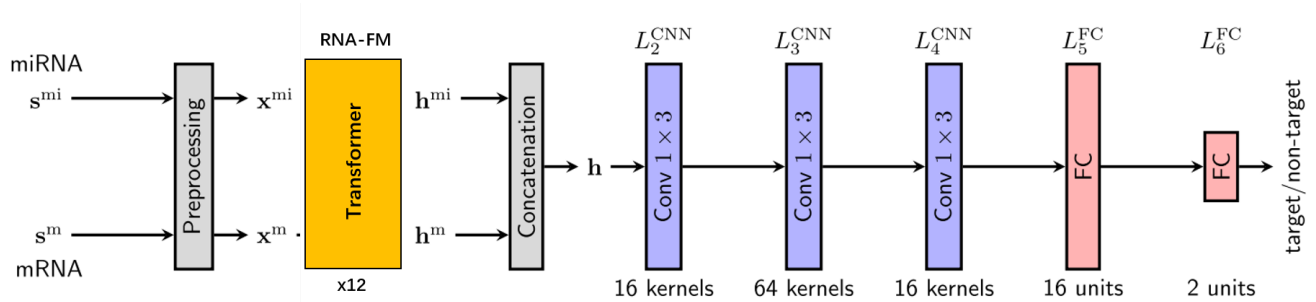
III Train & Inference

3.1 Train

Hyperparameter

- seed_match = 'offset-9-mer-m7'

- level = 'gene'

- **batch_size = 32**

- **epochs = 10**

```
Epoch 1/10   ████████████████████   2045/2045 [01:33<00:00, loss=0.267, acc=0.881]
Epoch 2/10   ████████████████████   2045/2045 [01:35<00:00, loss=0.221, acc=0.905]
Epoch 3/10   ████████████████████   2045/2045 [01:34<00:00, loss=0.209, acc=0.908]
Epoch 4/10   ████████████████████   2045/2045 [01:35<00:00, loss=0.203, acc=0.91]
Epoch 5/10   ████████████████████   2045/2045 [01:35<00:00, loss=0.198, acc=0.912]
Epoch 6/10   ████████████████████   2045/2045 [01:34<00:00, loss=0.195, acc=0.913]
Epoch 7/10   ████████████████████   2045/2045 [01:34<00:00, loss=0.192, acc=0.914]
Epoch 8/10   ████████████████████   2045/2045 [01:35<00:00, loss=0.19, acc=0.916]
Epoch 9/10   ████████████████████   2045/2045 [01:35<00:00, loss=0.187, acc=0.917]
Epoch 10/10  ████████████████████   2045/2045 [01:35<00:00, loss=0.185, acc=0.919]
```

3.2 Inference

Hyperparameter

- **seed_match = 'offset-9-mer-m7'**

- **level = 'gene'**

Result

DeepTarget

	accuracy
set 1	0.7729
set 2	0.7934
set 3	0.7948
set 4	0.7690
set 5	0.7707
set 6	0.7758
set 7	0.7787
set 8	0.7822
set 9	0.7753
set 10	0.7907
average	0.7804

Our model

```
                                                          156/156 [00:06<00:00]
4962
acc: 79.32285368802901 %
                                                          171/171 [00:07<00:00]
5457
acc: 84.91845336265347 %
                                                          162/162 [00:07<00:00]
5153
acc: 82.8449446924122 %
                                                          155/155 [00:06<00:00]
4956
acc: 81.79983857949959 %
                                                          160/160 [00:07<00:00]
5101
acc: 81.96432072142717 %
                                                          169/169 [00:07<00:00]
5388
acc: 79.28730512249443 %
                                                          160/160 [00:07<00:00]
5102
acc: 79.08663269306155 %
                                                          162/162 [00:07<00:00]
5167
acc: 82.11728275595122 %
                                                          156/156 [00:06<00:00]
4986
acc: 82.49097472924187 %
                                                          162/162 [00:07<00:00]
5172
acc: 83.33333333333334 %
avg =  81.71659396781038 %
```

Reference

1.1

miRNA: 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

3' - A C G G U G C C G G **G A G C G G G** U - 5'

seed region

∅ ∅ | | | | ∅ ∅ ∅ | | | | | | | | |

Offset 9-mer-m7

mRNA:

5' - G C A A C C A C A A U C C U C G C C A C U - 3'

candidate target site (CTS) of size k
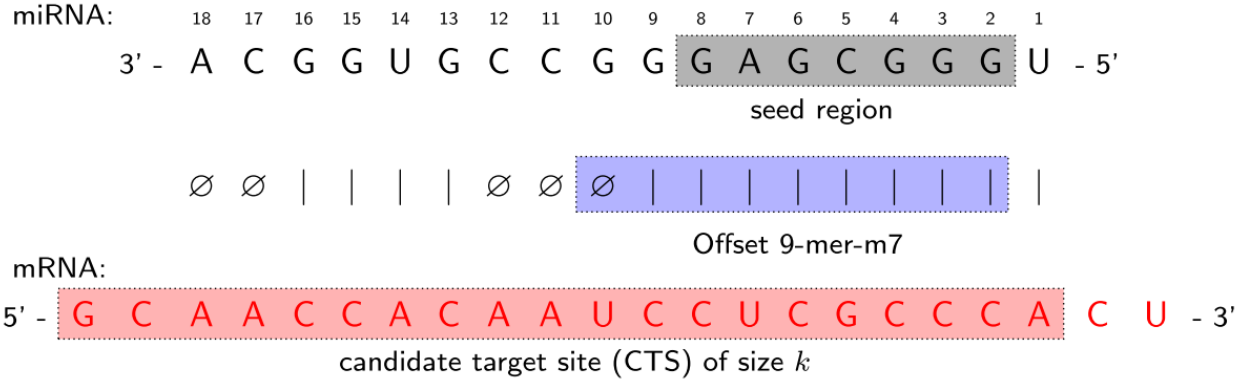
FIGURE 1. Definition of terminologies in interactions between miRNAs and mRNAs (best viewed in color). The gray, blue, and red box represent a seed region, a relaxed site pattern (e.g., Offset 9-mer-m7), and a CTS of size $k = 20$, respectively.

1.1.1 Define the CTSs

- **10-mer-m6**: six WC pairings from the miRNA nucleotides 1–10

- **10-mer-m7**: seven WC pairings from the miRNA nucleotides 1–10

- **Offset 9-mer-m7**: seven WC pairings from the miRNA nucleotides 2–10

1.2
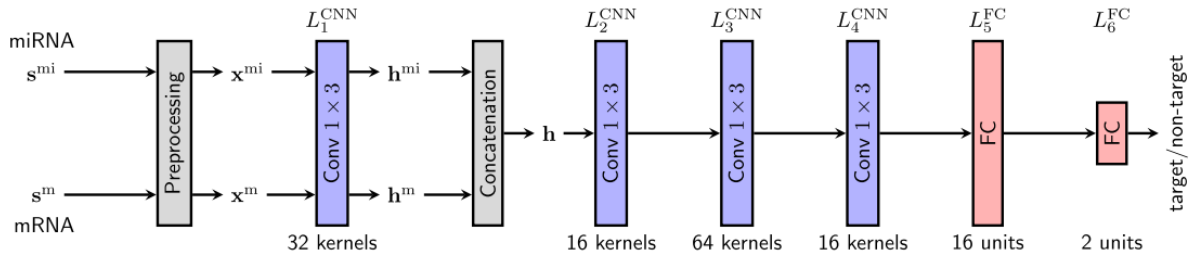
FIGURE 2. Overview of the proposed methodology.

Algorithm 1 MicroRNA Target Prediction

Input: N miRNA-mRNA pairs, $\mathbf{s}_1^{\text{mi}}, \cdots, \mathbf{s}_N^{\text{mi}}$ and $\mathbf{s}_1^{\text{m}}, \cdots, \mathbf{s}_N^{\text{m}}$

Output: $\mathbf{y}$ (target/non-target)

1: search CTSs of given miRNA-mRNA pairs (see Algorithm 2)   ▷ $\hat{\mathbf{s}}_1^{\text{mi}}, \cdots, \hat{\mathbf{s}}_N^{\text{mi}}$ and $\hat{\mathbf{s}}_1^{\text{m}}, \cdots, \hat{\mathbf{s}}_N^{\text{m}}$

2: encode miRNA and mRNA sequences in CTSs   ▷ $\mathbf{x}_1^{\text{mi}}, \cdots, \mathbf{x}_N^{\text{mi}}$ and $\mathbf{x}_1^{\text{m}}, \cdots, \mathbf{x}_N^{\text{m}}$

3: feedforward a shared [CNNs] layer   ▷ $[\cdot]$: a shared layer for both $\mathbf{x}^{\text{mi}}$'s and $\mathbf{x}^{\text{m}}$'s

4: concatenate two representations, $\mathbf{h}^{\text{mi}}$ and $\mathbf{h}^{\text{m}}$, into $\mathbf{h}$ ▷ $\mathbf{h}^{\text{mi}}, \mathbf{h}^{\text{m}}$: outputs of the [CNNs]

5: feedforward the additional stacked CNNs and FC layers

6: **repeat**

7:      minimize the logarithmic loss $\mathcal{L}_{\text{w}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}(w^1 y_i \log(p_i) + w^0(1 - y_i)\log(1 - p_i))$

8: **until** # of epoch is n_{epoch}

- shared CNNs layer for both miRNAs and mRNAs learns representations of both sequences

- additional stacked CNNs and fully-connected (FC) layers model sequence-to-sequence interactions between miRNA-mRNA pairs (lines 4–8)