

Micro RNA target prediction – A Deep Learning approach

Student ID: 200103251

Module: Final Project

Date: 17/07/2023

Template: Machine Learning and Neural Networks – Deep learning on a public dataset

1. Background

The central dogma of biology states that DNA is copied into mRNA (transcription), and the information encoded in the mRNA is used to synthesize proteins (translation) [8]. Since mRNAs carry the information required to produce proteins, the levels of a specific mRNA affect the levels of the respective protein it codes for [8]. Proteins dictate functions in the organism, hence their changes in concentration may lead to expressing phenotypes, diseases, and clinical conditions [8, 21].

At the DNA level, gene activity can also be determined based on the abundance of the produced RNA transcripts, where high amounts of transcript levels mean an upregulation of the gene, and low transcript quantities imply gene downregulation; similarly, protein levels can be inferred from the abundance of their respective transcript (mRNA). The abundance of synthesized RNA molecules present in a sample provides a snapshot of the status of the genes, constituting a valuable tool in determining which ones are on or off and assessing their activity [21].

Among all types of RNA, miRNAs constitute key components of the network of gene regulatory

pathways [10, 11], and act by pairing with imperfectly complementary mRNA strands to downregulate gene expression and modulate cell activity [2, 5]. Hence, miRNAs are promising as therapeutic agents, potentially overcoming the limitations of small drug molecules that target only certain proteins [13]. They can also overcome concerns associated with monoclonal antibodies which are highly specific but limited to circulating proteins and cell-surface receptors, because, unlike current therapeutic molecules, miRNAs can downregulate the expression of almost all genes/transcripts [13].

To perform its function, the miRNA must successfully pair with an mRNA target strand. The targeting process determines the effect that a miRNA will exert on gene expression levels, the respective coding protein, and consequently on biological functions. Despite its relevance in the mechanisms of multiple biological processes and disease states, the targeting process is poorly understood, and the current approaches require incorporating previous knowledge into traditional ML models [1]; however, some studies have proved that with the use of DL techniques on RNA data, more accurate predictions can be generated on large-scale datasets at the expense of interpretability and confidence on the decision-making process [1, 3, 5, 7].

2. Objectives

- Gather and prepare publicly available data about curated miRNA-mRNA target

interactions on the selected organism (i.e., *A.thaliana*).

- Build and train a DNN to predict mRNA targets for miRNA candidates.
- Identify factors likely to influence the targeting process of miRNA molecules (interpretability of the DNN and biological meaning of hidden layers).

3. Justification

In organisms, gene regulation constitutes a tool for responding to environmental changes by controlling gene expression alongside developmental and physiological processes, including reparation and disease mechanisms [15, 16]. Gene regulation is highly mediated by ncRNAs which are RNA sequences that do not code for a protein and whose functions remain mostly unknown. Among ncRNAs, miRNAs are fundamental for regulating gene expression and silencing pathways, even in the presence of imperfectly complementary gene targets; hence, to successfully characterize complex regulatory pathways in the organisms, there is an increasing need for understanding the mechanisms of action of miRNAs. [17]

Since the effect of the miRNA on the organism mainly depends on the functions of the protein encoded by the target mRNA molecule [2], for every identified miRNA, it is essential to identify the set of targeted molecules. Nonetheless, considering that miRNAs can pair with imperfectly complementary strands and the vast amount of unique mRNA sequences in an organism, performing experiments to identify the targets may result unviable in terms of resources [18]. Therefore, the most accepted approach consists of running in silico experiments to generate a set of candidates of

manageable size, which are then further confirmed either in vitro or in vivo.

Although classical ML approaches have been applied to the problem of identifying miRNA targets, constraints and previous knowledge are expected to be included as input for the algorithm; therefore, this approach represents a challenge due to the gaps in the literature and the lack of understanding about the overall process [1, 19].

Having identified the relevance and current limitations of miRNA target prediction, a DL approach is proposed as a suitable solution for discovering patterns in high-dimensional data. However, categorized as black box systems, DL approaches present challenges in terms of interpretability and reliability of the decision-making process; this concern is particularly relevant in the health industry and medical field, where a wrong decision can lead to fatal consequences for a given organism, patient, or population. [1, 3, 19]

Although the interpretability of DNN on RNA data is a field that has not been explored thoroughly, there are studies focusing on the implementation and interpretation of DNN aiming to extract the factors (e.g., genes, proteins, etc.) that exert the most influence on the predictions for a given clinical condition; however, these studies do not explain the neurons and have not explored the representation learned in the hidden layers of the network [19].

4. Scope

4.1. Domain

This project contributes to the field of bioinformatics and research on mechanisms of molecule interactions.

The selected data is generated from the organism *Arabidopsis thaliana*. Although *A. thaliana* is considered the key model for plant biology, it can be used to understand human diseases due to the conservation of protein function, conservation of cellular processes, and the high percentage of genes shared between both species [31, 32]. Furthermore, the insights produced may be useful in understanding the mechanisms of other species of plants and crops [32].

Such observations imply that this study overlaps with the agriculture domain but also relates to the health sector.

4.2. Users

The target public of this project comprises the scientific community and pharmaceutical companies, hence, this study is relevant not only at the academic level but also at the industrial level. In both cases (academic or industrial purposes), the users are either entities or individuals conducting research on miRNA-mRNA interactions.

4.3. Use cases

Academic:

- Contribute to the understanding of gene regulation mechanisms mediated by miRNAs.
- Assist in the functional annotation task for known and novel miRNAs.
- Provide the information required to update the biological pathways associated with the proteins coded by the target mRNAs.
- Propose novel hypotheses for diseases and conditions with unknown etiology.

Industry:

- Reduce the number of candidate miRNA-mRNA pairs to validate experimentally in drug discovery research.
- Assessing the potential risk of side effects due to interacting targets in clinical trials.

Note: This paper differentiates between academic and industry based on the following criteria: “Academic” implies no commercial value, while “Industry” represents commercial value.

5. Literature review

DL techniques are starting to be explored in the field of bioinformatics, with satisfactory results compared to traditional ML algorithms [1]. Although for RNA data most attempts have focused on exploring DL approaches on RNA sequencing and gene expression levels [3, 5, 7, 22], there are some studies aiming to solve the miRNA targeting problem [1, 23]. In terms of interpretability, there are few RNA sequencing publications exploring the explainability of the network [19], but no studies aiming for the interpretability of miRNA target prediction. This section presents and analyzes 4 of the most representative similar works.

1. *“Biological interpretation of deep neural network for phenotype prediction based on gene expression” [19]:*

This study focuses on RNA sequencing data to predict clinical conditions using a DNN. Given the high dimensionality of this type of data, for every patient, the gene expression levels were reduced using autoencoders. A DNN was trained, achieving better

performance when compared to traditional ML models (Figure 1). [19]

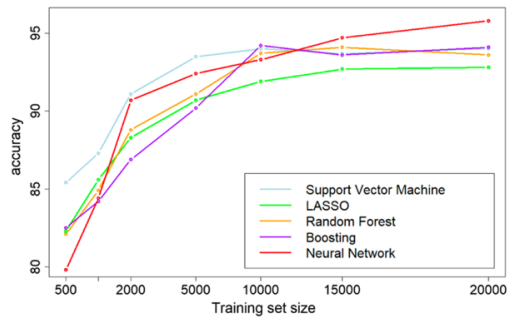


Figure 1. Accuracy of the different learning algorithms in the function of the training set size. [19]

Additionally, the authors identified that similar studies exploring the interpretability of the network were focused only on prediction interpretation (i.e., explain the prediction of a given input) rather than model interpretation (i.e., explain the logic of the model when predicting for the whole population); therefore, this study aimed to explore the model interpretability to investigate the representation of the gene expression learned in the hidden layers. [19]

The results proved that given a DL model that successfully finds relationships between gene expression levels and phenotypes, there should be a link between both variables and therefore, by exploring the interpretation, new biological hypotheses can be proposed to be experimentally investigated. [19]

This study was selected because it is a pioneer not only in the field of using DL on RNA sequencing data but also in DNN interpretation beyond the identification of impactful genes. There is also a strong ethical component motivating the researchers because of the high relevance of

the decisions that such a model could have when applied to real-world patients. Although this study is similar to the proposed for this project in terms of using DL techniques on RNA data and exploring the interpretability of the resulting model, it is focused on another type of RNA data (i.e., RNA sequencing vs. miRNA-mRNA target pairs) and has different goals (i.e., predict phenotypes vs. predict mRNA targets).

2. *“DeepMirTar: a deep-learning approach for predicting human miRNA targets” [23]*

In this study, the authors aimed to apply DL techniques to the problem of identifying binding sites of target mRNA sequences for miRNAs. The approach followed was to use a SdA, a type of NN consisting of multiple layers each one with massive units. The generated tool targets sites at the 3’UTR region and considers only the seed region of miRNAs, consisting of the first 8 nucleotides of the sequence and which is usually involved in the binding site. [23]

Another relevant consideration is the selection of the organism.; unlike other studies, this one proposes training the network only on Homo sapiens RNA data. The mechanisms of action by which the miRNA targets a mRNA strand may differ between organisms, thus constraining the model to consider a single organism could lead to better predictions. [23]

The resulting tool achieved higher performance when compared to state-of-the-art approaches., including DT (Decision Trees), LR (Logistic Regression), RF (Random Forests), MLP (Multilayer Perceptron), and CNN (Convolutional Neural Networks). [23]

This paper was selected because of the rigorous comparison the authors performed to evaluate DL and alternative ML approaches, which motivates the use of DL techniques applied to biological data. Both projects differ in terms of interpretability goals, and organism selection, and while the researchers considered only the miRNA seed region, the proposed project aims to consider the full miRNA length to account for non-canonical pairing.

3. *“miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts” [1]*

This study was a pioneer in the field of prediction of miRNA-mRNA target pairs by considering beyond the seed region of the miRNA sequences. Instead of restricting the model to work with miRNA seed regions, the researchers included more flexibility to include the entire miRNA and 3'UTR mRNA nucleotides. [1]

The data cited by the authors include only Homo sapiens sequences and was used to train a DNN composed of autoencoders and a feed-forward network. The results of the trained network consistently showed that DL approaches outperform traditional state-of-the-art ML algorithms and succeeded in recognizing the relevance of the seed region in the targeting process. Although the results were consistent with the literature in determining that the miRNA seed region plays an important role in the target selection, the network also identified pairs outside the canonical criteria. Such findings confirmed that in order to understand

miRNA processes, the whole strands should be considered. [1]

The relevance of the length of the considered miRNA sequence and the findings described in the study motivate the approach used in the project proposal. Hence, the proposed project and this study overlap in terms of topic and consideration of entire miRNA strands but differ in interpretability goals and target organism.

4. *“Interpretable drug target prediction using deep neural representation” [28]*

The authors of this paper aimed to propose a NN model to predict drug-target interactions using low-level representations as input. Alongside the predicted interactions, the model provided biological interpretation, which was not proposed in previous works. [28]

The most reliable and interpretable approach to generating predictions of interactions is molecular docking; however, this approach is often unaffordable in terms of time and resources and is limited by the availability of 3D protein structures. Considering the above, the authors modeled the problem as a binary classification task for an ML model able to receive drug-target pairs and predict if they will interact or not. [28]

In the proposed model, input pairs constitute low-level representations where the target is encoded as a raw amino acid sequence alongside GO (Gene Ontology) terms, and the drug is represented as either a chemical structure graph or as a SMILES sequence string. Drugs in the form of

sequential structures should be converted into chemical structure graphs that become dense vector representations and can be exploited by attention mechanisms; the conversion was carried out using LSTM RNN. [28]

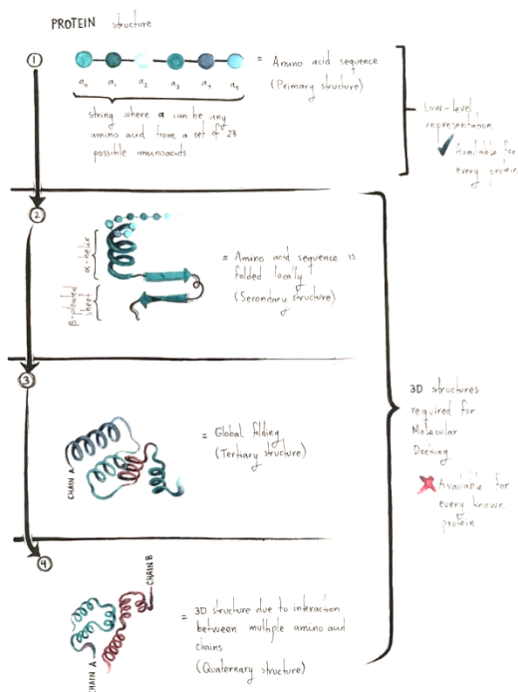


Figure 2. Diagram interpretation of the input representations for proteins [28], and how the decisions differ from traditional approaches.

In the case of drugs that are not sequential, no conversion is required, hence the inputs are processed directly by a CNN adapted to consider neighboring atoms instead of neighboring pixels; such CNN applies a filter to each atom and its neighbors, so it requires only 5 filters because the range of possible neighbors for a single atom is [1, 5]. As a result of applying the filter, the CNN captures local signals that are then aggregated and pooled to produce a final vector representation. [28]

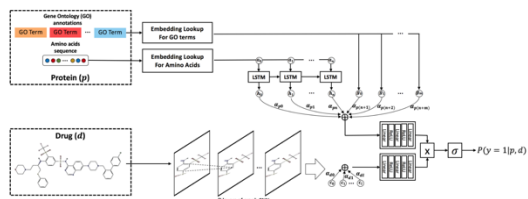


Figure 3. Diagram representation of the data flow and architecture of the NN. [28]

The next stage consists of an Attentive pooling network receiving an interaction as input and generating a matrix based on the interaction between every single amino acid in the target and every atom in the drug; the matrix undergoes then row-wise max-pooling operations to generate the attention weights for the drugs, and column-wise max pooling to generate the attention weights for the targets. The attention weights are normalized by SoftMax to get the attention-based vector representations. [28]

Once the attention-based vector representations are calculated, they are used at the inference stage to feed a Siamese network consisting of two multilayer networks. Each vector representation goes to one of the networks and the two respective outputs are operated using the inner product; the Sigmoid function is applied to the result which generates a value representing the probability of interaction. A threshold for the classification boundary is set, and the prediction is formulated based on it. The result was an approach able to generalize to new proteins while providing biological insights to understand the prediction process. [28]

6. Design

This Machine Learning and Neural Networks project is structured to answer if a miRNA-mRNA pair is likely to interact (derived from *A. thaliana*), which constitutes the main research question. Secondary research questions can be derived from the main one, including which mechanisms are involved in the pairing process, or to what extent the results can be extrapolated to other organisms.

To answer the research question, this proposal includes a methodology with 4 main stages. The first stage consists of building and training the Deep Neural Network on experimentally confirmed interacting *A. thaliana* miRNA-mRNA pairs. The second phase is related to the interpretability of the network, which will be explored through attention layers.

For the DNN stage, the following architecture is proposed:

- **RNN (Recurrent Neural Network) – LSTM** (Long Short Term Memory)

The project will use a LSTM network, which is a type of RNN. This decision relies on the fact that, although the model will accept sequential representations (strings of nucleotides), RNA molecules have 3-dimensional structures (similar to the protein structure described in Figure 3). Since the 3-dimensional nature can be categorized as a long-distance dependency [28], LSTM and their ability to memorize long-term data will be valuable.

This stage should take as input the sequential data representing RNA sequences (either micro-RNA or messenger

RNA). The inputs should be converted into dense vector representations that can be exploited with attention in later steps. The proposed DNN using LSTM aims to produce such representations and output the respective hidden vectors. From this step, the hidden space for each molecule should be inferable, and a matrix of the interaction of the pair is expected as one of the outputs. In this project, such an interaction matrix would consist of a matrix stating the interaction between each nucleotide from the input sequences.

- **2-way attention – Attentive Pooling Network**

This step relies on 2-way attention networks to allow both inputs from the miRNA-mRNA pair to be aware of each other. The interaction matrix generated in the previous phase should be the input for this stage. Using the interaction matrix, context matrices should be generated for both molecules. Then, the interaction matrix can undergo row-wise and column-wise Max-pooling operations to extract the attention weights for miRNA and mRNA respectively.

The output of this stage is the set of attention weights for each molecule.

- **Softmax**

The next stage consists of taking the attention weights and normalizing them. For this purpose, the Softmax function is proposed. The outputs of this step are the attention-based vector representations (one per RNA molecule).

- **Inference**

The last stage is based on a Siamese network [28] composed of 2 independent input networks. The network takes as input the attention-based vector representations and they are fed separately. Then both outputs are operated using the inner product, followed by the Sigmoid function. The result after applying the Sigmoid function is a probability of an interaction between the molecules. This probability can be used alongside a defined threshold to use as a classification boundary [28].

The following diagram summarizes the main 4 stages of the proposed project.

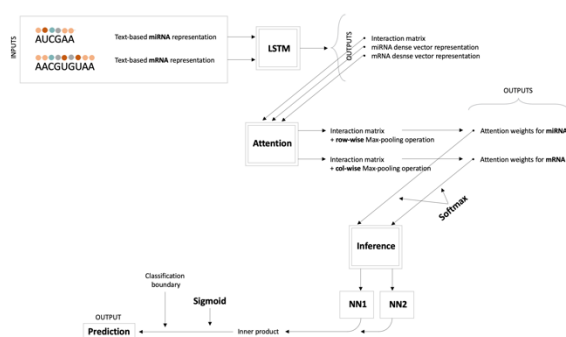


Figure 4. Diagram representation of the proposed architecture.

7. Dataset

The data for this project was extracted from miRTarBase, a database containing experimentally confirmed interactions between miRNAs and mRNAs from the same organism [14]. Considering the scope of this study, the interactions of the organism *A. thaliana* were collected.

The dataset includes the sequential representation of the molecules, and the method of validation used experimentally.

The dataset is available at https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/download.php.

8. Evaluation strategy

The dataset will be divided into training, validation, and testing datasets. However, the separation of the datasets should not be performed without ensuring the presence and balance of negative (non-interacting pairs) and positive (interacting pairs) data. This constitutes a challenge given the type of data, where only a few experiments have proved negative interactions. Since it could be possible for the datasets to lack negative examples, in the training stage, the sampling process will be stratified to ensure the presence of both classes [1].

Another concern to be addressed in terms of evaluation is the probability of overlapping data in the training and testing sets. This can happen because of miRNA families, impairing the ability of the network to generalize [1]. To overcome this problem, those overlapping miRNAs will be excluded.

The network will be trained and validated following k-fold cross-validation. Prediction scores and ROC curves will be evaluated, and their significance will be assessed using a Wilcoxon signed rank test [1].

Additionally, the results will be compared against the current gold standards in miRNA target prediction [1]:

- TargetScan
- Diana microT-CDS
- PITA
- miRanda
- mirzaG
- Paccmit
- mirDB

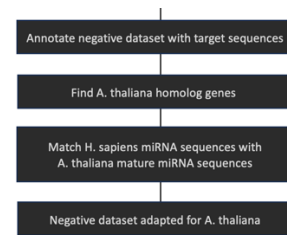


Figure 5. Methodology diagram.

Implementation

Since this project presents a challenge in terms of negative data availability, the earlier stages of the implementation are focused on generating a negative dataset (i.e., miRNA and mRNA combinations that do not interact). For this purpose, a curated negative dataset for H. sapiens [33] is used to match homology sequences in the selected organism A. thaliana.

The mature miRNA sequences for both organisms are retrieved from miRBase [36], and the sequences of the target mRNA molecules are extracted from the respective genomes [29, 37, 38].

```

In 17: # Load the FASTA file containing all known miRNA sequences for all organisms.
1 with open('data/mature_miRNA_all_organisms.fa') as f:
2     mature_miRNAs = f.read().split('\n')[1:]
3     f.close()
4
5 # Isolate H. sapiens (has) and A. thaliana (ath) sequences.
6 hsa_mature_miRNAs_dict = {miRNA.split('.')[0]: miRNA.split('.')[1]}
7     for miRNA in mature_miRNAs:
8         if 'has' in miRNA:
9             continue
10        if 'ath' in miRNA:
11            ath_mature_miRNAs_dict[miRNA.split('.')[0]] = miRNA.split('.')[1]
12        for miRNA in mature_miRNAs:
13            if 'ath' in miRNA:
14                continue
15            if 'has' in miRNA:
16                hsa_mature_miRNAs_dict[miRNA.split('.')[0]] = miRNA.split('.')[1]
17
18 print('Total has miRNAs: ', len(hsa_mature_miRNAs_dict))
19 print('Total ath miRNAs: ', len(ath_mature_miRNAs_dict))
20
21 Executed at 2023-08-18 20:23:05 in 50ms
22
23 Total has miRNAs: 2555
24 Total ath miRNAs: 359
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529

```

References

- [1] A. Pla, X. Zhong, and S. Rayner, "miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts," *PLOS Computational Biology*, vol. 14, no. 7, p. e1006185, Jul. 2018, doi: 10.1371/journal.pcbi.1006185.
- [2] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng, "Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation," *Frontiers in Endocrinology*, vol. 9, 2018, Accessed: May 04, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fendo.2018.00402>
- [3] A. Quillet et al., "Improving Bioinformatics Prediction of microRNA Targets by Ranks Aggregation," *Frontiers in Genetics*, vol. 10, 2020, Accessed: May 04, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fgene.2019.01330>
- [4] H. Nakayashiki, 'RNA silencing in fungi: Mechanisms and applications', *FEBS Letters*, vol. 579, no. 26, pp. 5950–5957, Oct. 2005, doi: 10.1016/j.febslet.2005.08.016.
- [5] T. Kakati, D. K. Bhattacharyya, J. K. Kalita, and T. M. Norden-Krichmar, 'DEGnext: classification of differentially expressed genes from RNA-seq data using a convolutional neural network with transfer learning', *BMC Bioinformatics*, vol. 23, no. 1, p. 17, Jan. 2022, doi: 10.1186/s12859-021-04527-4.
- [6] B. Hanczar, F. Zehraoui, T. Issa, and M. Arles, 'Biological interpretation of deep neural network for phenotype prediction based on gene expression', *BMC Bioinformatics*, vol. 21, no. 1, p. 501, Nov. 2020, doi: 10.1186/s12859-020-03836-4.
- [7] D. Urda, J. Montes-Torres, F. Moreno, L. Franco, and J. M. Jerez, 'Deep Learning to Analyze RNA-Seq Gene Expression Data', in *Advances in Computational Intelligence*, I. Rojas, G. Joya, and A. Catala, Eds., in *Lecture Notes in Computer Science*, vol. 10306. Cham: Springer International Publishing, 2017, pp. 50–59. doi: 10.1007/978-3-319-59147-6_5.
- [8] 'Central Dogma', *Genome.gov*, Sep. 14, 2022. <https://www.genome.gov/genetics-glossary/Central-Dogma> (accessed May 07, 2023).
- [9] A. Talukder, W. Zhang, X. Li, and H. Hu, "A deep learning method for miRNA/isomiR target detection," *Sci Rep*, vol. 12, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41598-022-14890-8.
- [10] O. P. Gupta, P. Sharma, R. K. Gupta, and I. Sharma, "Current status on role of miRNAs during plant–fungus interaction," *Physiological and Molecular Plant Pathology*, vol. 85, pp. 1–7, Jan. 2014, doi: 10.1016/j.pmpp.2013.10.002.
- [11] E. Marín-González and P. Suárez-López, "'And yet it moves': Cell-to-cell and long-distance signaling by plant microRNAs," *Plant Science*, vol. 196, pp. 18–30, Nov. 2012, doi: 10.1016/j.plantsci.2012.07.009.
- [12] T. Siddika and I. U. Heinemann, "Bringing MicroRNAs to Light: Methods for MicroRNA Quantification and Visualization in Live Cells," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2021, Accessed: Apr. 18, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fbioe.2020.619583>
- [13] J. K. W. Lam, M. Y. T. Chow, Y. Zhang, and S. W. S. Leung, "siRNA Versus miRNA as

Therapeutics for Gene Silencing,” *Mol Ther Nucleic Acids*, vol. 4, no. 9, p. e252, Sep. 2015, doi: 10.1038/mtna.2015.23.

[14] “miRTarBase: the experimentally validated microRNA-target interactions database.” https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php (accessed May 08, 2023).

[15] “Gene Regulation,” *Genome.gov*, Sep. 14, 2022. <https://www.genome.gov/genetics-glossary/Gene-Regulation> (accessed May 09, 2023).

[16] C. Stylianopoulou, “Carbohydrates: Regulation of metabolism,” in *Encyclopedia of Human Nutrition* (Fourth Edition), B. Caballero, Ed., Oxford: Academic Press, 2023, pp. 126–135. doi: 10.1016/B978-0-12-821848-8.00173-6.

[17] L. He and G. J. Hannon, “MicroRNAs: small RNAs with a big role in gene regulation,” *Nat Rev Genet*, vol. 5, no. 7, Art. no. 7, Jul. 2004, doi: 10.1038/nrg1379.

[18] D. Pradhan, A. Kumar, H. Singh, and U. Agrawal, “Chapter 4 - High-throughput sequencing,” in *Data Processing Handbook for Complex Biological Data Sources*, G. Misra, Ed., Academic Press, 2019, pp. 39–52. doi: 10.1016/B978-0-12-816548-5.00004-6.

[19] B. Hanczar, F. Zehraoui, T. Issa, and M. Arles, “Biological interpretation of deep neural network for phenotype prediction based on gene expression,” *BMC Bioinformatics*, vol. 21, no. 1, p. 501, Nov. 2020, doi: 10.1186/s12859-020-03836-4.

[20] A. L. Leitão and F. J. Enguita, “A Structural View of miRNA Biogenesis and Function,” *Non-Coding RNA*, vol. 8, no. 1, Art. no. 1, Feb. 2022, doi: 10.3390/ncrna8010010.

[21] ‘Gene Expression | Learn Science at Scitable’.

<https://www.nature.com/scitable/topicpage/gene-expression-14121669/> (accessed May 07, 2023).

[22] W. Guo, Y. Xu, and X. Feng, ‘DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing’. *arXiv*, May 08, 2017. doi: 10.48550/arXiv.1705.03094.

[23] M. Wen, P. Cong, Z. Zhang, H. Lu, and T. Li, ‘DeepMirTar: a deep-learning approach for predicting human miRNA targets’, *Bioinformatics*, vol. 34, no. 22, pp. 3781–3787, Nov. 2018, doi: 10.1093/bioinformatics/bty424.

[24] X. M. Xu and S. G. Møller, ‘The value of Arabidopsis research in understanding human disease states’, *Curr Opin Biotechnol*, vol. 22, no. 2, pp. 300–307, Apr. 2011, doi: 10.1016/j.copbio.2010.11.007.

[25] G. P. Way and C. S. Greene, ‘Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders’. *bioRxiv*, p. 174474, Aug. 11, 2017. doi: 10.1101/174474.

[26] J. Rocca, ‘Understanding Variational Autoencoders (VAEs)’, *Medium*, Mar. 21, 2021. <https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73> (accessed Jun. 07, 2023).

[27] C. H. Grønbech, M. F. Vording, P. Timshel, C. K. Sønderby, T. H. Pers, and O. Winther, ‘scVAE: Variational auto-encoders for single-cell gene expression data’. *bioRxiv*, p. 318295, Oct. 02, 2019. doi: 10.1101/318295.

[28] K. Y. Gao, A. Fokoue, H. Luo, A. Iyengar, S. Dey, and P. Zhang, ‘Interpretable Drug Target Prediction Using Deep Neural Representation’, in

Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 3371–3377. doi: 10.24963/ijcai.2018/468.

[29] 'Arabidopsis thaliana (ID 4) - Genome - NCBI'.
https://www.ncbi.nlm.nih.gov/genome/4?genome_assembly_id=380024 (accessed Jul. 02, 2023).

[30] G. B. Or and I. Veksler-Lublinsky, 'Comprehensive machine-learning-based analysis of microRNA-target interactions reveals variable transferability of interaction rules across species'. *bioRxiv*, p. 2021.03.28.437385, Mar. 29, 2021. doi: 10.1101/2021.03.28.437385.

[31] 'Arabidopsis thaliana (ID 4) - Genome - NCBI'.
https://www.ncbi.nlm.nih.gov/genome/4?genome_assembly_id=380024 (accessed Jul. 02, 2023).

[32] X. Chen, 'Small RNAs – secrets and surprises of the genome', *Plant J*, vol. 61, no. 6, pp. 941–958, Mar. 2010, doi: 10.1111/j.1365-3113X.2009.04089.x.

[33] S. Bandyopadhyay and R. Mitra, 'TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples', *Bioinformatics*, vol. 25, no. 20, pp. 2625–2631, Oct. 2009, doi: 10.1093/bioinformatics/btp503.

[34] 'PmiREN: Plant microRNA Encyclopedia'.
<https://www.pmiren.com/download> (accessed Aug. 04, 2023).

[35] 'refSeq Accession to Gene Symbol Converter - Genomics Biotools'.

https://www.biotools.fr/mouse/refseq_symbol_converter (accessed Aug. 07, 2023).

[36] 'miRBase - Downloads'.
<https://mirbase.org/download/> (accessed Aug. 13, 2023).

[37] 'Genome', NCBI.
<https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=9606> (accessed Aug. 13, 2023).

[38] '11968211 - Assembly - NCBI'.
https://www.ncbi.nlm.nih.gov/assembly/?term=GCF_000001405 (accessed Aug. 13, 2023).

[39] B. Murcott, R. J. Pawluk, A. V. Protasio, R. Y. Akinmusola, D. Lastik, and V. L. Hunt, 'stepRNA: Identification of Dicer cleavage signatures and passenger strand lengths in small RNA sequences', *Frontiers in Bioinformatics*, vol. 2, 2022, Accessed: Aug. 18, 2023. [Online].