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.

5.1. “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]

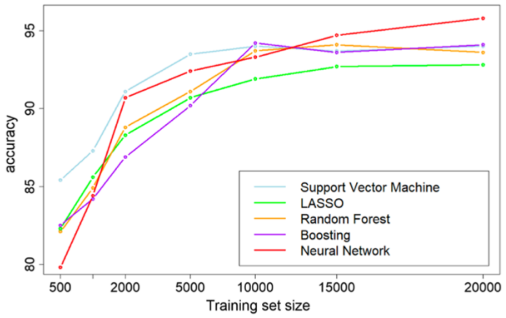


Figure 1. Accuracy of the different learning algorithms in the function of the training set size. [19]

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).

5.2. *“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.

5.3. *"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.

5.4. *"Interpretable drug target prediction using deep neural representation" [28]*

The authors of this paper aimed to propose an 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]

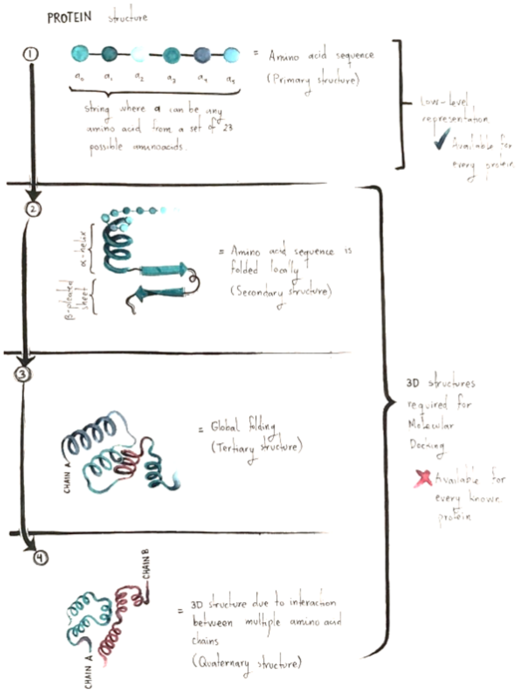

PROTEIN structure
1
a_0 a_1 a_2 a_3 a_4 a_5
= Amino acid sequence (Primary structure)
string where a can be any amino acid from a set of 23 possible aminoacids.
Low-level representation
✓ Available for every protein
2
α-helix
β-pleated sheet
= Amino acid sequence is folded locally (Secondary structure)
3
CHAIN A
= Global folding (Tertiary structure)
3D structures required for Molecular Docking
✗ Available for every known protein
4
CHAIN B
CHAIN A
= 3D structure due to interaction between multiple amino acid chains (Quaternary structure)

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]

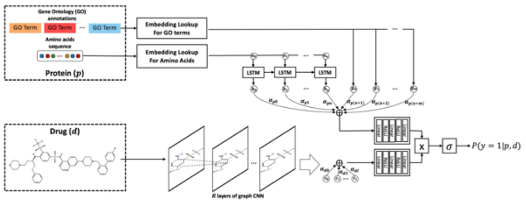


Figure 3. Diagram representation of the data flow and architecture of the NN. [28]

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.

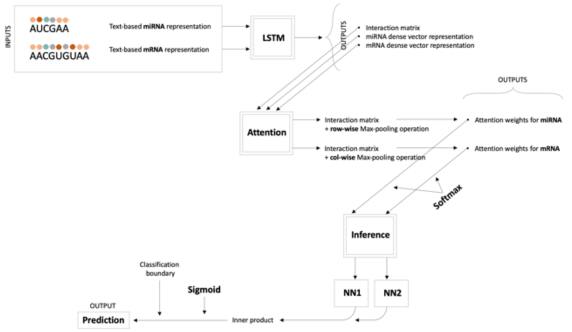

INPUTS
AUCGAA
AACGUGUAA
Text-based **miRNA** representation
Text-based **mRNA** representation
LSTM
OUTPUTS
Interaction matrix
miRNA dense vector representation
mRNA desnse vector representation
OUTPUTS
Attention
Interaction matrix
+ **row-wise** Max-pooling operation
Interaction matrix
+ **col-wise** Max-pooling operation
Attention weights for **miRNA**
Attention weights for **mRNA**
Softmax
Inference
NN1
NN2
Classification boundary
Sigmoid
OUTPUT
Prediction
Inner product

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].

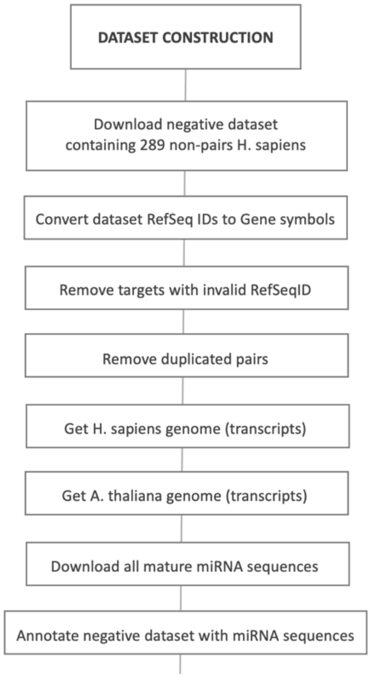
Additionally, the results will be compared against the current gold standards in miRNA target prediction [1]:

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

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].


DATASET CONSTRUCTION
Download negative dataset containing 289 non-pairs H. sapiens
Convert dataset RefSeq IDs to Gene symbols
Remove targets with invalid RefSeqID
Remove duplicated pairs
Get H. sapiens genome (transcripts)
Get A. thaliana genome (transcripts)
Download all mature miRNA sequences
Annotate negative dataset with miRNA sequences

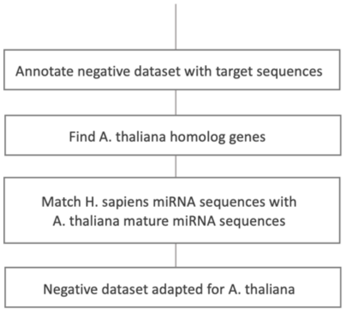
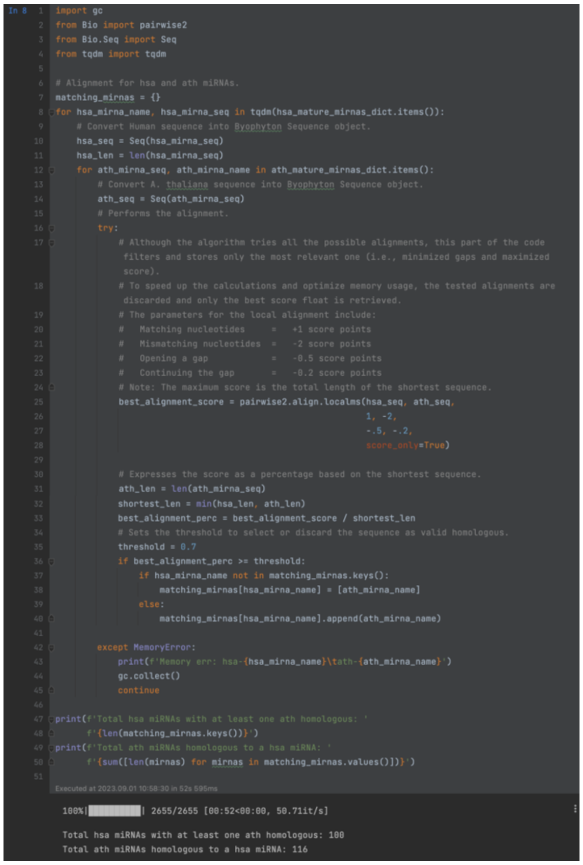

Annotate negative dataset with target sequences
Find A. thaliana homolog genes
Match H. sapiens miRNA sequences with A. thaliana mature miRNA sequences
Negative dataset adapted for A. thaliana

Figure 5. Methodology diagram.

The objective of this stage is to generate a negative dataset based on the experimentally validated non-target interactions



```
import gc
from Bio import pairwise2
from Bio.Seq import Seq
from tqdm import tqdm

# Alignment for hsa and ath miRNAs.
matching_mirnas = {}
for hsa_mirna_name, hsa_mirna_seq in tqdm(hsa_mature_mirnas_dict.items()):
    # Convert Human sequence into Byophyton Sequence object.
    hsa_seq = Seq(hsa_mirna_seq)
    hsa_len = len(hsa_mirna_seq)
    for ath_mirna_seq, ath_mirna_name in ath_mature_mirnas_dict.items():
        # Convert A. thaliana sequence into Byophyton Sequence object.
        ath_seq = Seq(ath_mirna_seq)
        # Performs the alignment.
        try:
            # Although the algorithm tries all the possible alignments, this part of the code
             filters and stores only the most relevant one (i.e., minimized gaps and maximized
             score).
            # To speed up the calculations and optimize memory usage, the tested alignments are
             discarded and only the best score float is retrieved.
            # The parameters for the local alignment include:
            #   Matching nucleotides     =   +1 score points
            #   Mismatching nucleotides  =   -2 score points
            #   Opening a gap            =   -0.5 score points
            #   Continuing the gap       =   -0.2 score points
            # Note: The maximum score is the total length of the shortest sequence.
            best_alignment_score = pairwise2.align.localms(hsa_seq, ath_seq,
                                                           1, -2,
                                                           -.5, -.2,
                                                           score_only=True)

            # Expresses the score as a percentage based on the shortest sequence.
            ath_len = len(ath_mirna_seq)
            shortest_len = min(hsa_len, ath_len)
            best_alignment_perc = best_alignment_score / shortest_len
            # Sets the threshold to select or discard the sequence as valid homologous.
            threshold = 0.7
            if best_alignment_perc >= threshold:
                if hsa_mirna_name not in matching_mirnas.keys():
                    matching_mirnas[hsa_mirna_name] = [ath_mirna_name]
                else:
                    matching_mirnas[hsa_mirna_name].append(ath_mirna_name)

        except MemoryError:
            print(f'Memory err: hsa-{hsa_mirna_name}\tath-{ath_mirna_name}')
            gc.collect()
            continue

print(f'Total hsa miRNAs with at least one ath homologous: '
      f'{len(matching_mirnas.keys())}')
print(f'Total ath miRNAs homologous to a hsa miRNA: '
      f'{sum([len(mirnas) for mirnas in matching_mirnas.values()])}')
```

Executed at 2023.09.01 10:58:30 in 52s 595ms

```
100%|██████████| 2655/2655 [00:52<00:00, 50.71it/s]

Total hsa miRNAs with at least one ath homologous: 100
Total ath miRNAs homologous to a hsa miRNA: 116
```

Figure 6. Code implementation. Construction of a negative dataset based on Homo sapiens non-pair instances. MiRNA matching phase.

Figure 7. Code implementation. Sequence matching and homology search on A. thaliana.

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.1365-313X.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].

[40] H.-Y. Huang et al., 'miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions', Nucleic Acids Research, vol. 50, no. D1, pp. D222–D230, Jan. 2022, doi: 10.1093/nar/gkab1079.

[41] 'Bio.pairwise2 module — Biopython 1.75 documentation'. https://biopython.org/docs/1.75/api/Bio.pairwise2.html (accessed Aug. 29, 2023).

[42] 'miRBase'. https://www.mirbase.org/ftp.shtml (accessed May 08, 2023).

[43] 'TAIR - Home Page'. https://www.arabidopsis.org/index.jsp (accessed Sep. 07, 2023).

[44] 'Home - Nucleotide - NCBI'. https://www.ncbi.nlm.nih.gov/nuccore/ (accessed Sep. 20, 2023).