

Beyond Sequence: A Novel Image-Based Model for MicroRNA Target Prediction

1st Mohammad Mohebbi

Computer Science and Info. Science
University of North Georgia
Dahlonega, United States
mmohebbi@ung.edu

2nd Spencer Sesser

Computer Science and Info. Science
University of North Georgia
Dahlonega, United States
sesess9594@ung.edu

3rd Phillip Williams

Computer Science and Info. Science
University of North Georgia
Dahlonega, United States
phillip.williams@ung.edu

4th Maurice Wodtke

Computer Science and Info. Science
University of North Georgia
Dahlonega, United States
mewodt4050@ung.edu

5th Sabastian Highton

Computer Science and Info. Science
University of North Georgia
Dahlonega, United States
sthgh7790@ung.edu

6th Shehenaz Shaik

Department of Computing
East Tennessee State University
Johnson City, United States
shaiks@etsu.edu

Abstract—MicroRNAs are small non-coding RNAs that regulate gene expressions in plants and animals. They bind to a gene and stop the process of protein synthesis from the targeted gene. They, through this process, control cell activities such as proliferation, differentiation, and death rate.

To understand the functionality of a microRNA, we need to know the targets of the microRNA. Often biologists do in vivo experiments to identify microRNA's targets. However, these experiments are expensive, time-consuming, and limited.

In the last two decades, bioinformaticians have developed countless computational models that predict a microRNA's target based on sequence information. But these models mainly suffer from high false positive rates and mediocre recall rates, due to incomplete understanding of the mechanism of microRNA targeting.

In this paper, we propose the Base Pair Probability Image (BPPI) model; a novel approach that converts a microRNA target prediction question into an image detection problem. We create an image for each pair of microRNA and target (or non-target) sequences, where the pixel values represent the probabilities of interactions between nucleotides from the two sequences. Then, we use a deep learning model to detect images depicting a microRNA and its target. Our model outperforms several widely used microRNA target prediction software tools, such as miRanda and RNAhybrid. This work has the potential to be used by biologists to shortlist potential microRNA targets for further verification through in vivo experiments.

Index Terms—Artificial Intelligence, Predictive Model, MicroRNA, Machine Learning, Image Classification

I. INTRODUCTION

MicroRNAs are tiny RNA sequences with variable lengths between 19 to 24 nucleotides [1]. They regulate gene expression through two main mechanisms: 1) binding to the 3'-untranslated region (3'-UTR) of mRNA sequences and stopping the ribosome translation. This process is called gene silencing or gene downregulation [2]. 2) Eliminating an mRNA that is no longer needed or when it has abnormal features. The latter one is used in bacteria [3].

MicroRNAs regulate several cell activities such as developmental timing, cell differentiation, death rate, etc. Dysfunction

of microRNAs could cause various diseases, including cancer, neurodegenerative disorders [4], and an increasing number of genetic diseases [5]. MicroRNAs' ability to regulate gene expression makes them promising tools for cancer treatment and drug design. MicroRNAs can serve as biomarkers for early cancer detection; their expression patterns in various cancers offer crucial insights into the stage and aggressiveness of the disease, facilitating personalized treatment planning [6]. Currently, it is estimated that the human body has roughly 2200 microRNAs and around 20,000 to 25,000 genes, but only a small fraction of genes is expressed in a given cell. MicroRNAs are believed to regulate about 30% of gene expressions in humans. [7] Given the great potential of microRNAs and their significant roles in cellular activities and gene expression, researchers are eager to know the functionality of microRNAs. Discovering the functionality of each microRNA is challenging for several reasons: 1) The same microRNA can target different mRNAs in different cell types or under different conditions. 2) MicroRNAs can interact with other microRNAs or proteins and affect the expression of a gene that they do not necessarily bind to. [8] 3) Current microRNA target prediction algorithms often have high false positive rates. [9] 4) Experimental methods of target discovery or target validation are slow and costly.

In this work, we introduce a new Artificial Intelligence (AI) based and data-driven method for accurate prediction of human microRNA targets. This method consists of two novel components: We create an image of possible interactions between a microRNA and its target, and then we train a Deep Neural Network (DNN) on the images of microRNAs and their experimentally validated targets. We demonstrate that these approaches make it possible to predict microRNA targets with a performance exceeding the state-of-the-art methods by a large margin.

II. DATA SET

The effectiveness of data-driven techniques is highly dependent on the data's quality. To create the most precise models and the most realistic assessments, we gathered our data from mirTarBase [10], one of the most up-to-date datasets and the most frequently cited resource for microRNA target prediction research. Notably, mirTarBase contains over 360,000 experimentally verified microRNA-target pairs from 18 different species. We are primarily interested in building and evaluating our machine learning method using human data.

We extracted human microRNA and 3'-UTR of mRNA gene sequences with experimentally known structures from mirTarBase. MirTarBase has a huge number of computationally discovered target genes. We only extracted experimentally validated samples to make our model to be completely clear from being biased toward the computational methods used by other researchers.

In mirTarBase, for each pair of a microRNA and the 3'-UTR of the targeted mRNA gene, there is a specific location, called the target site, that the microRNA binds to. We chose the target site as a positive sample for the microRNA and we fixed its length to 25 nucleotides. However, negative samples were not readily available in mirTarBase. In theory, any sequence of appropriate length within the 3'-UTR of a targeted mRNA gene that is not the actual target site can be considered a negative target for the corresponding microRNA. For each positive sample, we randomly selected ten locations within the 3'-UTR of a targeted mRNA gene to obtain negative samples at a ratio of ten to one. Each sample consists of a 22-nucleotide microRNA sequence and a 25-nucleotide site sequence that represents the actual target site for positive samples or a non-target site for negative samples.

A. Test set and Training set

For our experiments we divided our data into two sets; a training set (80%) and a test set (20%). The human data we extracted from mirTarBase contained 322 distinct microRNAs, 3651 target site sequences, and 3722 pairs of microRNAs and target sites. On average, each microRNA had more than ten target sites. Selecting test set samples randomly from the entire database would likely result in a significant overlap of microRNAs between the test and training sets. To prevent such overlaps and ensure the independence of the test set, we indexed microRNA-target site pairs based on microRNA sequences. Additionally, to create a test set with a distribution similar to that of the entire dataset, we sorted samples by microRNA sequence. We then included four consecutive microRNA sequences (based on the sorting order) and all their target and non-target sites in the training set, followed by one microRNA sequence and all its targets and non-targets in the test set, and so on. This approach ensured that the test set and training set had no overlaps in terms of microRNA sequences and that the test set's distribution closely resembled that of the entire dataset. Both test sets and training sets maintained a 1:10 ratio of positive to negative samples. The test set comprised 6127 samples (557 positives and 5570 negatives).

III. MODEL AND METHODS

A. Computing probability of canonical base pairs in RNA molecules

RNA sequences are composed of four nucleotides: adenine (A), cytosine (C), guanine (G), and uracil (U). These nucleotides, which for the rest of the paper we demonstrate as A, C, G, and U, bind to each other and form base pairs. The most common types of base pairs so-called canonical base pairs are AU, CG, GU, and vice versa. To compute the probability of different base pairs we went through the following procedure: 1) We extracted ribosomal RNA or rRNA molecule structures as Protein Data Bank (PDB) files from RCSB.org [11]. The reason we chose rRNAs is because they are a class of long non coding RNAs that share intriguing connections with microRNAs in regulating gene expression [12]. Moreover, they are well studied and their structure data is widely available [11, 13], while the structural bindings of microRNAs to mRNAs remain poorly understood, with limited data on their interactions. The data we collected contains 139 various-sized and single-chain molecules. 2) As shown in Figure 1, we used x3DNA-DSSR software [14] to extract the secondary structure of rRNA molecules. Secondary structures are composed of canonical base pairs. None canonical base pairs are not provided in secondary structures. We calculated their frequencies and computed the probability of different base pairs. The results were $P(\text{AU or UA}) = 0.21$, $P(\text{CG or GC}) = 0.67$, and $P(\text{GU or UG}) = 0.12$.

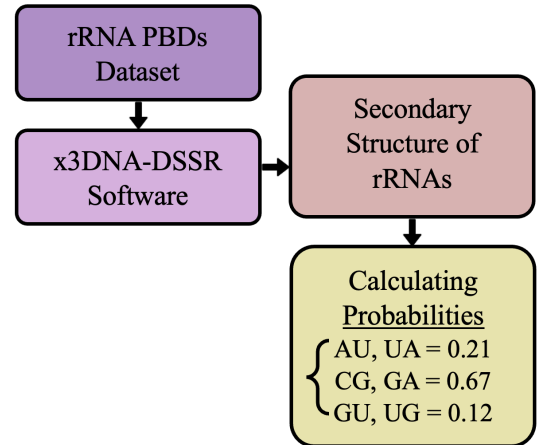


Fig. 1. We extracted 139 single chain and various size rRNA molecules PDB files from RCSB.org and used x3DNA-DSSR to extract the secondary structure of molecules. We counted base pairs in the structures and computed the probabilities of different canonical base pairs as shown in the figure.

B. Creating image of a microRNA and its target interactions

For each pair of a microRNA sequence and target (or non-target) sequence, we create a probabilistic image. The image dimensions are *length of microRNA* by *length of target*, i.e. 22 by 25 pixels. The value of the pixel at row i and column j is equal to the probability of base pairing between nucleotides at index i of the microRNA sequence and at index j of the target (or non-target) sequence. The process is shown in Figure 2.

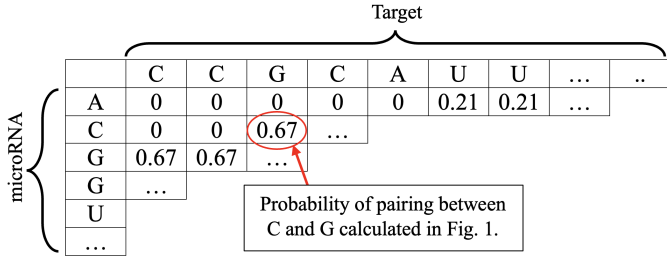


Fig. 2. We create an image for a pair of microRNA and target candidate sequences. Each image pixel value is the probability of a base pair between the corresponding nucleotides of the two sequences.

IV. BASE PAIR PROBABILITY IMAGE (BPPI) MODEL

At the core of our method lies a Deep Neural Network (DNN) that detects images of microRNAs and their targets. The image is converted into a one dimensional vector of values by appending its rows. This vector then serves as the input for the DNN model; the output, or label, is 1 if the image depicts a microRNA and its target, and is 0 if it shows a microRNA and a non-target.

Our DNN comprises four layers: input, output, and two hidden layers. The number of nodes in the input and two hidden layers equals the size of the input vector, which is $22 \times 25 = 550$. The output layer has one node for the label. The input layer *kernel_initializer* parameter is set to 'uniform', and activation function in layers one to three is 'relu', and the last layer's activation function is 'sigmoid'.

The process of converting microRNA and target sequences to images and train a DNN on the images is shown in Figure 3. We trained DNN with epoch parameter 5 and batch size value 100.

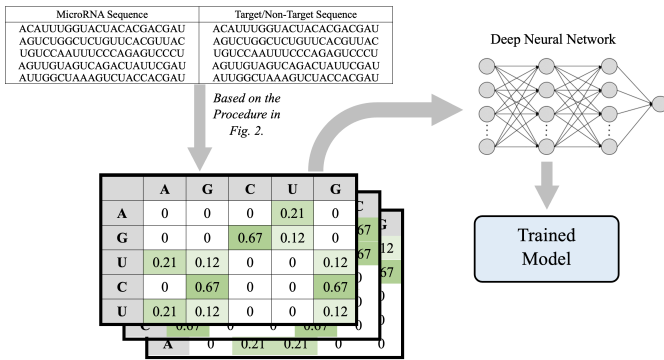


Fig. 3. BPPI Model: For our dataset of microRNA and target candidate sequences, for each pair of a microRNA and a target, we create images as shown in Fig. 2. We build and train a DNN for the dataset of images. For a new pair of microRNA and target candidate, or an unseen sample, we follow the same procedure to create their image, then the trained DNN predicts if the candidate sequence is a target for the microRNA.

V. RESULTS

To have a comprehensive comparison of our model versus other methods, we selected two bioinformatics tools that are widely used for microRNA target prediction and two machine

learning algorithms; RandomForest and LinearRegression. The two state-of-the-art bioinformatics target prediction algorithms are miRanda[15] and RNAhybrid[16]. We selected these methods based on the criteria that are mentioned in the Discussion section.

We measured the performance of the aforementioned methods and ours in terms of precision, recall and specificity. Two common plots to compare the performance of a binary classifier, which distinguish positive samples from negatives, are Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves. For each curve, a higher line indicates a superior performance for the classifier.

Figure 4 shows PR and ROC curves for our model versus the other methods. In the figure, our model is labeled as BPPI. BPPI outperforms other methods by a clear margin in both metrics. This is largely due to the fact that we incorporated rRNA base pair probabilities into an image of a microRNA and its target pair. The closest competitors to BPPI are RandomForest and RNAhybrid. Random Forest classifier is a powerful machine learning algorithm that excels at tackling classification problems. It has been widely used in gene expression analysis, and we were not surprised that it came second in the competition. Third place goes to RNAhybrid, a powerful microRNA target prediction method designed by bioinformaticians and validated by biologists. RNAhybrid uses a variety of biological rules governing sequence-to-sequence bindings, such as sequence complementarity, structural motifs, and the thermodynamics of microRNA and gene binding, to detect targets. Linear Regression model performs the worst but still outperforms a random baseline classifier. This indicates that the majority of interactions between microRNAs and targets cannot be accurately modeled by a linear function.

Bioinformatics software miRanda has a mediocre performance and it is because this algorithm uses factors such as the number and location of base pairs and the Minimum Free Energy (MFE) of a target and microRNA binding. There are several studies suggesting that the binding between microRNA and target may not necessarily be the least free energy binding location in a targeted gene [17].

We believe that, for our analysis, the PR curve is a more appropriate comparison method, because in our test set the ratio of positives to negatives is 1 to 10, and PR curve is robust even in case of very imbalanced test sets.

To quantitatively compare methods with PR and ROC curves, researchers often compute the area under them. The area under the PR curve is denoted by AUPRC and the area under the ROC curve is abbreviated as AUC. A perfect classifier has an AUC and AUPRC of 1. A baseline random classifier has $AUC = 0.5$ but its AUPRC depends on the ratio of positive samples to the total number of samples; $AUPRC = (\text{Number of positive samples}) / (\text{Total number of samples})$. In our test sets, we have one positive sample per ten negative samples; therefore, a random classifier has $AUPRC = 1/11 = 0.09$. Table I compares the performance of BPPI with the other methods based on AUPRC and AUC. BPPI with AUPRC of 0.71 and AUC of 0.92 outperforms other methods again.

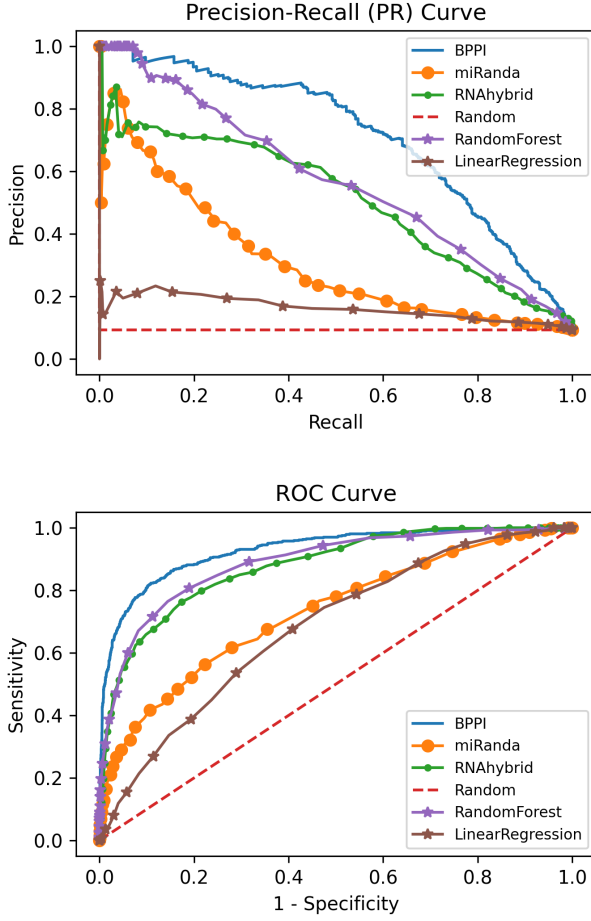


Fig. 4. The PR and ROC curves visualize the performance of our model BPPI to both bioinformatics target prediction algorithms like miRanda and RNAhybrid, and widely used classifiers like RandomForest and LinearRegression algorithms.

One of the issues with computational methods for microRNA target prediction is having a high false positive rate. For a binary classifier false positive rate and recall ratio depends on the threshold value we choose to classify samples either as positive or negative. As shown in table II, with Threshold value 0.11 we achieved a decent balance of false positive and recall percentages, 10% and 80% respectively. For the other methods we chose two threshold values; one to have a false positive rate close to that of BPPI, i.e. 10%, and the other to achieve a recall percentage close to 80%. The results in the table show that in the other methods cannot have both low false positive and high recall percentages at the same time.

VI. DISCUSSION

Over the past 20 years, many algorithms have been developed to predict microRNA targets. These algorithms vary in

TABLE I
COMPARISON OF BPPI AND OTHER METHODS BY AREA UNDER PR AND ROC CURVES.

	AUPRC	AUC
BPPI	0.71	0.92
miRanda	0.31	0.73
RNAhybrid	0.51	0.88
RandomForest	0.57	0.88
LinearRegression	0.16	0.68
Random classifier	0.09	0.50

TABLE II
COMPARISON OF BPPI AND OTHER METHODS BY FALSE POSITIVE AND RECALL PERCENTAGES.

	Threshold	False Positive	Recall
BPPI	0.11	10%	80%
miRanda	0.58	11%	42%
	0.41	53%	80%
RNAhybrid	0.54	10%	66%
	0.47	22%	80%
RandomForest	0.17	11%	71%
	0.15	19%	81%
LinearRegression	0.17	11%	27%
	0.09	54%	79%

the information they use, their accessibility, whether they are based on rules or data, and their underlying principles. To compare our method fairly with the best existing methods, we studied several well-known tools, such as TargetScan [18], TarPmiR [19], miRBase [20], DIANA-microT [21], miRanda, miTarget [22], RNAhybrid, Avishka [23], TargetSpy [24], miRWalk [25], and miRanda-mirSVR [26].

Several of these algorithms incorporate additional information beyond microRNA and target sequences, such as sequence conservation across species, conserved or non-conserved microRNA family, and miRBase annotation. This makes comparing our model, which relies solely on sequence information, to some methods technically impractical due to missing data. Additionally, some algorithms lack publicly available source code or executable files, making them accessible only through web interfaces, which are unsuitable for submitting large test sets like ours. Furthermore, some methods differ fundamentally in their functionality. For example, TarPmiR identifies multiple targets for a single microRNA across an entire mRNA, while our model focuses on prediction of binding between a specific microRNA and target site sequence. Considering these factors, we chose miRanda and RNAhybrid algorithms. These rule-based methods rely on established principles of microRNA targeting, such as quantity and the location of base pairs, minimum free energy of the binding, and stability of the secondary structure, which have proven to be reliable performance metrics. As a result, miRanda and RNAhybrid remain widely used, both independently and as core components of other algorithms.

VII. CONCLUSION

MicroRNAs are small non-coding RNA sequences that play a crucial role in gene expression in a variety of organisms, such

as plants, animals, and viruses. These sequences participate in important cellular activities, including developmental timing, cell differentiation, and death rate. MicroRNAs function by targeting and binding to specific genes and regulating the protein synthesis of the genes. Biologists want to know the function of each microRNA, which requires identifying its targets. However, the experimental process of finding microRNA targets is costly and time-consuming.

This paper presents a novel method of microRNA target prediction. It creates an image for each pair of microRNA and a target candidate. The image dimensions correspond to the lengths of the microRNA and target candidate. Pixel values represent the probabilities of base pairing between nucleotides of the microRNA and the target candidate, calculated from a large dataset of rRNA molecules. We chose rRNAs because they are one of the most conserved RNA families and represent the most abundant type of RNA. Notably, they are essential for protein synthesis and share functional connections with microRNAs.

We trained a Deep Learning model on these images to detect a microRNA's targets from non-targets. To train and test our Deep Learning model, we collected microRNAs and their targeted genes from mirTarBase, the most cited database of experimentally validated microRNA targets.

Our results demonstrate the effectiveness of this approach, as our model significantly outperforms existing target prediction tools. While most current tools compromise either on false positive rate or recall percentage, our method offers an 80% recall rate with a very low false positive rate of 10%. In addition, our work shows that many microRNA and target pairs do not exhibit the least free energy bindings. This finding explains the mediocre performance of some energy-based algorithms like miRanda on diverse and experimentally validated datasets.

Our model empowers biologists to explore a broader range of candidate targets for microRNAs before resorting to *in vivo* verification.

REFERENCES

- [1] M. Giurgiu, R. Kaltenbach, F. Ahrend, S. Weeks, H. Clifton, M. Boulido, V. Voloshin, J. Zhong, S. Harden, and A. Kofman, "Multiple genetic polymorphisms within microRNA targets and homologous microRNA-binding sites: Two more factors influencing microRNA-mediated regulation of gene expression," 2023.
- [2] S. Khan, H. Ayub, T. Khan, and F. Wahid, "MicroRNA biogenesis, gene silencing mechanisms and role in breast, ovarian and prostate cancer," *Biochimie*, vol. 167, pp. 12–24, 2019.
- [3] R. Shang, S. Lee, G. Senavirathne, and E. C. Lai, "microRNAs in action: biogenesis, function and regulation," *Nature Reviews Genetics*, pp. 1–18, 2023.
- [4] S. Li, Z. Lei, and T. Sun, "The role of microRNAs in neurodegenerative diseases: A review," *Cell Biology and Toxicology*, vol. 39, no. 1, pp. 53–83, 2023.
- [5] C. A. Jużwik, S. S. Drake, Y. Zhang, N. Paradis-Isler, A. Sylvester, A. Amar-Zifkin, C. Douglas, B. Morquette, C. S. Moore, and A. E. Fournier, "microRNA dysregulation in neurodegenerative diseases: A systematic review," *Progress in neurobiology*, vol. 182, p. 101664, 2019.
- [6] L. J. Galvão-Lima, A. H. Morais, R. A. Valentim, and E. J. Barreto, "miRNAs as biomarkers for early cancer detection and their application in the development of new diagnostic tools," *Biomedical engineering online*, vol. 20, pp. 1–20, 2021.
- [7] S. Komal, J.-J. Yin, S.-H. Wang, C.-Z. Huang, H.-L. Tao, J.-Z. Dong, S.-N. Han, and L.-R. Zhang, "MicroRNAs: Emerging biomarkers for atrial fibrillation," *Journal of Cardiology*, vol. 74, no. 6, pp. 475–482, 2019.
- [8] C. E. Condrat, D. C. Thompson, M. G. Barbu, O. L. Bugnar, A. Boboc, D. Cretoiu, N. Suci, S. M. Cretoiu, and S. C. Voinea, "miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis," *Cells*, vol. 9, no. 2, p. 276, 2020.
- [9] K. T. Arif, R. K. Okolicsanyi, L. M. Haupt, and L. R. Griffiths, "MicroRNA–target identification: A combinatorial in silico approach," in *MicroRNA Detection and Target Identification: Methods and Protocols*. Springer, 2023, pp. 215–230.
- [10] H.-Y. Huang, Y.-C.-D. Lin, S. Cui, Y. Huang, Y. Tang, J. Xu, J. Bao, Y. Li, J. Wen, H. Zuo *et al.*, "mirTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions," *Nucleic acids research*, vol. 50, no. D1, pp. D222–D230, 2022.
- [11] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, H. Chao, L. Chen, P. A. Craig, G. V. Crichlow, K. Dalenberg, J. M. Duarte *et al.*, "RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning," *Nucleic acids research*, vol. 51, no. D1, pp. D488–D508, 2023.
- [12] M. A. McCool, C. J. Bryant, and S. J. Baserga, "MicroRNAs and long non-coding RNAs as novel regulators of ribosome biogenesis," *Biochemical Society Transactions*, vol. 48, no. 2, pp. 595–612, 2020.
- [13] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic acids research*, vol. 49, no. D1, p. D10, 2021.
- [14] A. V. Colasanti, X.-J. Lu, and W. K. Olson, "Analyzing and building nucleic acid structures with 3DNA," *JoVE (Journal of Visualized Experiments)*, no. 74, p. e4401, 2013.
- [15] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander, "The microRNA.org resource: targets and expression," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D149–D153, 2008.
- [16] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, *Fast and effective prediction of mi-*

- croRNA/target duplexes*. Cold Spring Harbor Lab, 2004, vol. 10.
- [17] J. Ding, X. Li, and H. Hu, “MicroRNA modules prefer to bind weak and unconventional target sites,” *Bioinformatics*, vol. 31, no. 9, pp. 1366–1374, 2015.
 - [18] B. P. Lewis, C. B. Burge, and D. P. Bartel, “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets,” *cell*, vol. 120, no. 1, pp. 15–20, 2005.
 - [19] J. Ding, X. Li, and H. Hu, “Tarpmir: a new approach for microRNA target site prediction,” *Bioinformatics*, vol. 32, no. 18, pp. 2768–2775, 2016.
 - [20] S. Griffiths-Jones, R. J. Grocock, S. Van Dongen, A. Bateman, and A. J. Enright, “mirbase: microRNA sequences, targets and gene nomenclature,” *Nucleic acids research*, vol. 34, no. suppl_1, pp. D140–D144, 2006.
 - [21] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis *et al.*, “Diana-microt web server: elucidating microRNA functions through target prediction,” *Nucleic acids research*, vol. 37, no. suppl_2, pp. W273–W276, 2009.
 - [22] S.-K. Kim, J.-W. Nam, J.-K. Rhee, W.-J. Lee, and B.-T. Zhang, “mitarget: microRNA target gene prediction using a support vector machine,” *BMC bioinformatics*, vol. 7, no. 1, p. 411, 2006.
 - [23] A. Ghoshal, A. Grama, S. Bagchi, and S. Chaterji, “An ensemble svm model for the accurate prediction of non-canonical microRNA targets,” in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, 2015, pp. 403–412.
 - [24] M. Sturm, M. Hackenberg, D. Langenberger, and D. Frishman, “Targetspy: a supervised machine learning approach for microRNA target prediction,” *BMC bioinformatics*, vol. 11, no. 1, p. 1, 2010.
 - [25] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, “mirwalk–database: prediction of possible mirna binding sites by “walking” the genes of three genomes,” *Journal of biomedical informatics*, vol. 44, no. 5, pp. 839–847, 2011.
 - [26] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie, “Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites,” *Genome biology*, vol. 11, no. 8, p. R90, 2010.