Building Profiles for miRNA Target Prediction

A Senior Honors Thesis

Submitted in Partial Fulfillment of the Requirements
for Graduation in the Honors College

By

Lucas Galbier

Biology & Mathematics Major

The College at Brockport
May 9, 2017

Thesis Director: Dr. Rongkun Shen, Associate Professor, Biology

Abstract

MicroRNAs are very short non-coding RNAs. Since microRNAs play important roles in many biological process, the research of microRNAs is a burgeoning field with much promise. Due to the high cost of experimental approaches, many computational techniques and algorithms have been implemented to study microRNAs. However, current methods for determining the targets for miRNAs are far from accurate. To address this issue, we developed algorithms that produced profiles of miRNA recognition elements and features such binding energy threshold and conservation score. These profiles will be used to train a machine learning algorithm for miRNA target prediction.

Introduction

MicroRNAs (miRNAs) are about 21-22 nucleotide long, single-strand, non-coding RNA molecules that are naturally expressed and play important roles in posttranscriptional regulation. After initial production in the nucleus and subsequent exportation to the cytosol via exportin 5, miRNAs down-regulate the translation of their targeted messenger RNAs (mRNAs) by binding to the 3' Untranslated Region (3'UTR). The formation of the miRNA-Induced Silencing Complex (RISC) inhibits ribosomal activity and promotes the degradation of the mRNA poly-A tail. (**Figure 1**) Each miRNA might bind to hundreds of mRNA targets and each mRNA target might have multiple miRNA recognition elements, also known as MREs (**Figure 2**). Experimental methods to identify miRNA targets are high in cost. To address this, computer algorithms have been developed to predict miRNA targets (Cheng et al. 2016, Gomes et al. 2013). These algorithms predict unknown data

by implementing machine learning, a specialized artificial intelligence approach that guides the model to learn critical information from training data.

In this project, we generated profiles for miRNA target prediction for both training and testing of a machine learning model. We built these profiles using our unique high quality datasets of miRNA direct targets from RISCtrap (experimentally determined targets). (Cambronne et al. 2012) The profiles contain features of energy thresholds assessment for complementary matches between miRNA and MRE, conservation assessment, and structural accessibility estimation. The free binding energy is used to measure the strength of the MRE binding, and those with low binding energy will be filtered out. The conservation scores of the MREs provide a measurement of similarity between genomes that could help research on different species.

Methods

Current methods for determining the targets for known miRNAs are inaccurate and costly. A machine learning approach to MRE target prediction would be the best direction for our research. To implement such an approach, we developed and implemented an algorithm using the Python package for Linux to find the MREs based on the human miRNA extended seed sequences from miRBase v20. (miRBase) The extended seed sequences are 9-nucleotide long regions that bind to the 3'UTR of the mRNA. The complementary region on the mRNA is referred to as the MRE. The mRNAs used in this study were from the human hg19 RefSeq Genes. The matching of MRE and miRNA seed sequences were allowed to have some flexibility to allow some minimal mismatches, G:U wobble pairs, or

bulges, however, bugles were not accounted for in our study. (Reczko et al. 2012) Binding categories were recorded based on these irregularities. (**Figure 3**) The matched MRE and miRNA extended seed sequences were used to calculate the free binding energy using RNAhybrid software. (Rehmsmeier et al. 2004; Krüger et al. 2006) After filtering out the sites with low binding energy, the remaining MRE will be incorporated into the profiles (Reczko et al. 2016). We calculated the conservation scores of the MREs across 46 vertebrate genomes, since studies show that the MREs are conserved among various species (Reczko et al. 2012). The 3'UTR of mRNA data from human hg19 RefSeq Genes and human miRNA data from miRBase v20 was the input for the algorithm, and the resulting data was outputted to a separate file as a collection of profiles. Subsequent algorithms used these profiles as input. All of these features will be combined and used as input for the machine learning model for both training and prediction. This machine learning model will be developed via Pylearn2 using our custom dataset, a SoftMax regression model, and a stochastic gradient descent algorithm.

Results

The primary result of our work was the development of the algorithms in order to build the profiles for a future machine learning model. The first step in generating the profiles was to make a MRE search algorithm. Inputs for this algorithm were human miRNA sequences from miRBase v20 and mRNA sequences from human hg19 RefSeq genes. To search for patterns of similar strings, rather than exact strings, we implemented an inexact search approach using regular expressions. Regular expression is a powerful, built-in search tool that scans lines of text for subtexts using characters that generalize the search. It is

especially useful to search sequences with mismatches or insertions/deletions. When searching for different binding categories, the command called would be structured depending on the category being searched. As an example, suppose the miRNA extended seed sequence was ACTTACGGG. When searching for 9mer MREs, the regular expression would be structured as [TAGAATGCCC], the complementary sequence of the miRNA extended seed. When searching for 8mer MREs, the regular expression would be structured as [AGAATGCCC]. Once a match is found, one base prior to this sequence is also recorded and analyzed to make sure it fits the criteria for 8mer. This process was repeated for each binding category in particular order, so as to prevent double-counting. When a match was found and categorized, the IDs of both miRNA and mRNA strands, the extended seed sequence, the binding category, the MRE sequence, and the position of the MRE on the chromosome are added to the output text file. **Figure 4** outlines the pseudo-code for the entire algorithm.

The next algorithm determined the binding energy of the MRE sequence. The miRNA and MRE sequences of each profile were used as input for RNAhybrid. (Rehmsmeier et al. 2004; Krüger et al. 2006) The output of RNAhybrid included the binding energy threshold calculation and the approximate p-value of extreme value distribution, which were both appended to the end of each profile and added to an output text file. **Figure 5** outlines the pseudo-code for the algorithm.

Once the initial profiles were generated, we wrote another Python script to calculate the conservation scores of the MRE sequences. Conservation scores were obtained from

phastCons 46-way vertebrate multiple-alignment from UCSC Genome Browser. The conservation scores are listed in order of position on their respective chromosome, and by using the starting position recorded in each profile, the conservation score is summed from the 9 consecutive scores. **Figure 6** outlines the pseudo-code for the algorithm.

After running the algorithms that we developed, 1,796,319 profiles were generated, covering 1223 unique mRNA target sequences and 878 unique miRNA extended seed sequences. **Table 1** records the resulting category abundance.

Also included in this study are the high-confidence target datasets from RISCtrap, including experimentally determined targets for miR-124, and miR-132. (Cambronne et al. 2012) These datasets will be used for training of the Deep Learning Neural Network.

Future work

The resulting profiles built from the current work need to be trimmed based on relevant energy thresholds. A Deep Learning Neural Network will be developed as a Multilayer Perceptron model using the Pylearn2 package for Linux. Then, the machine learning model will be implemented, which will accommodate our profiles as input. (**Figure 7**) The model will be optimized using various datasets. Once shown to be reliable, this prediction tool will be made publicly available online to benefit the miRNA research community.

Acknowledgements

I'd like to thank my advisor, Dr. Rongkun Shen, and my parents for providing feedback and critique on this paper, as well as the encouragement of my friends.

Figures and Tables

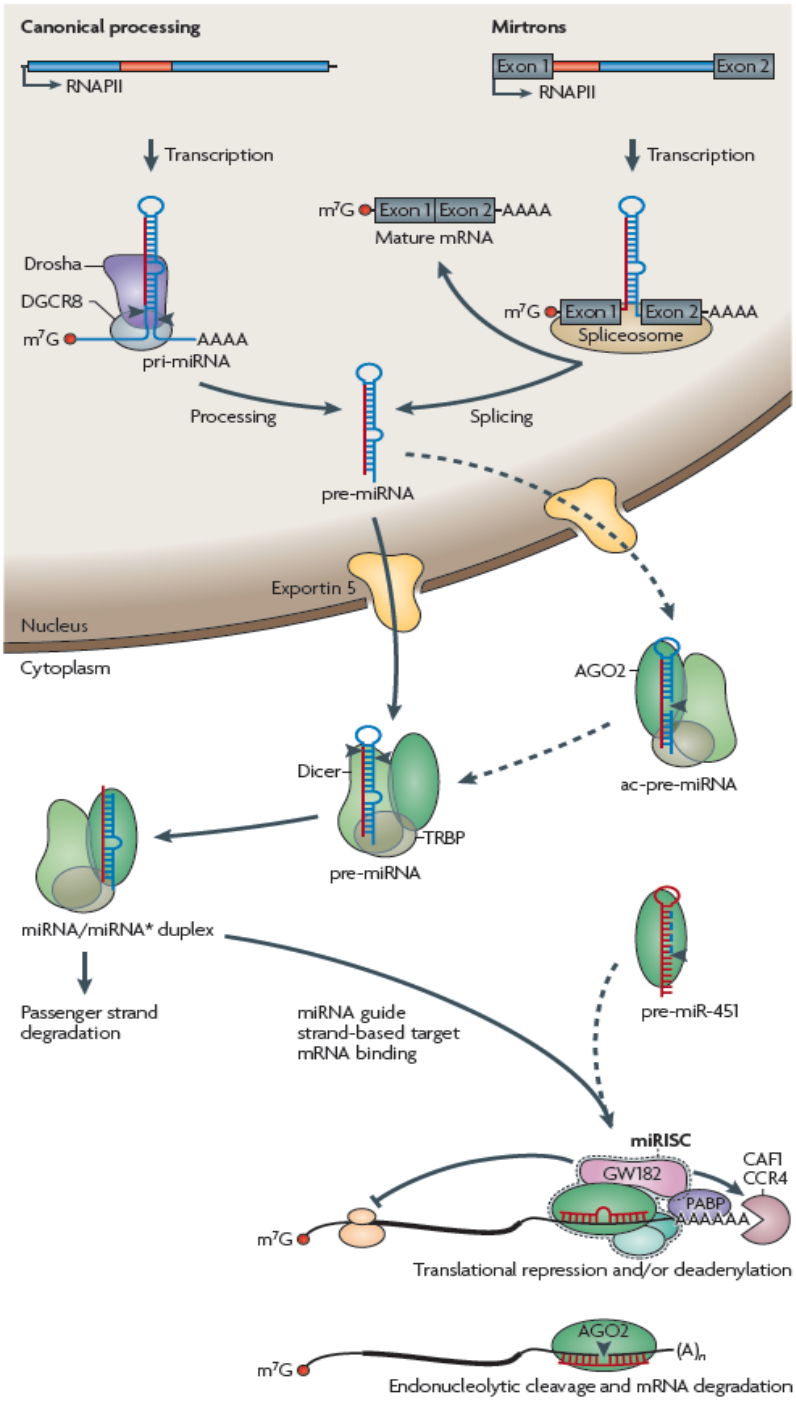


Figure 1. The biochemical pathway of miRNA production. (Krol, et al. 2010)

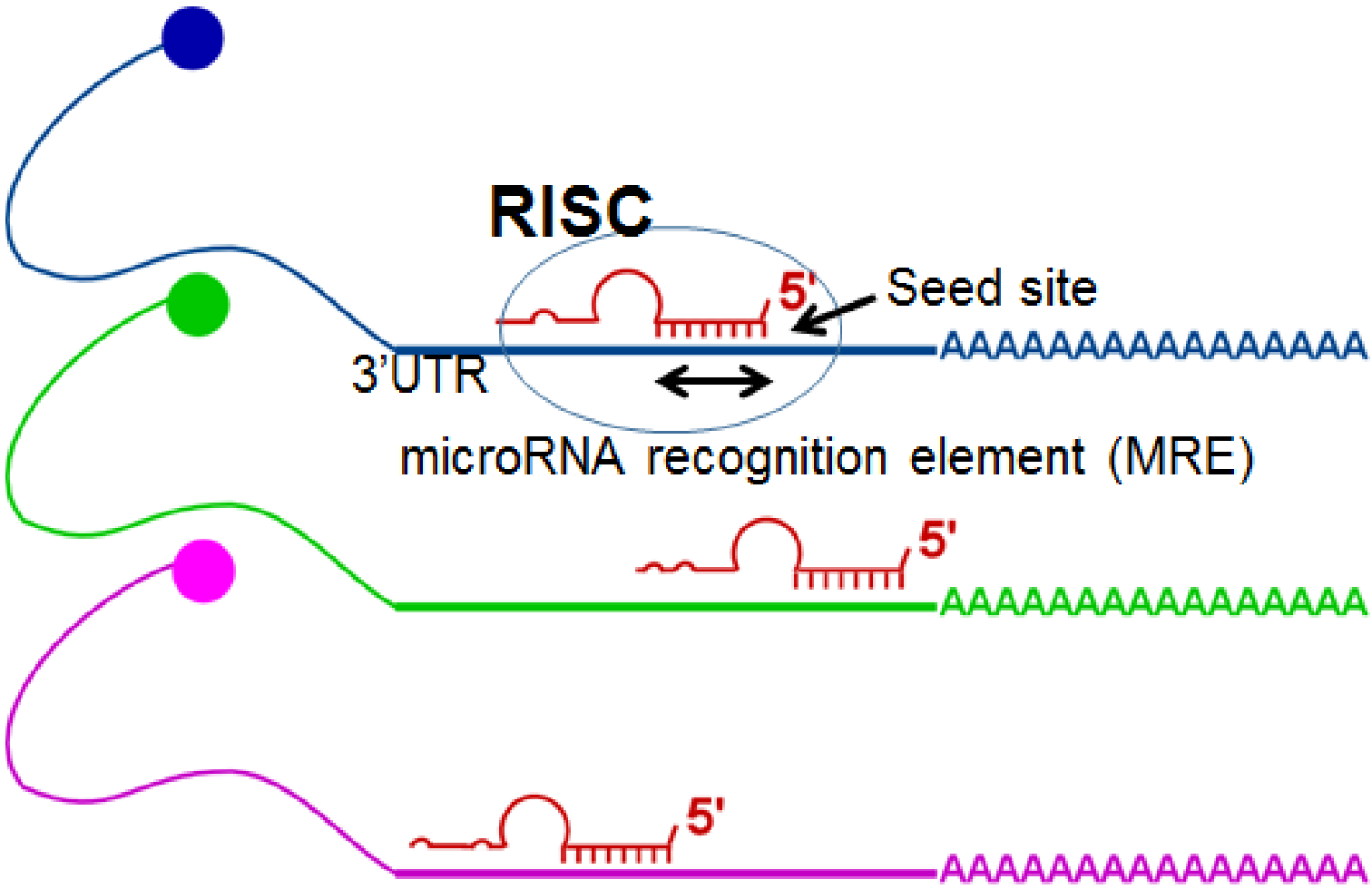


Figure 2. The interaction of miRNAs with the 3'UTR of target mRNAs. Extended seeds and MREs are identified, where binding occurs.

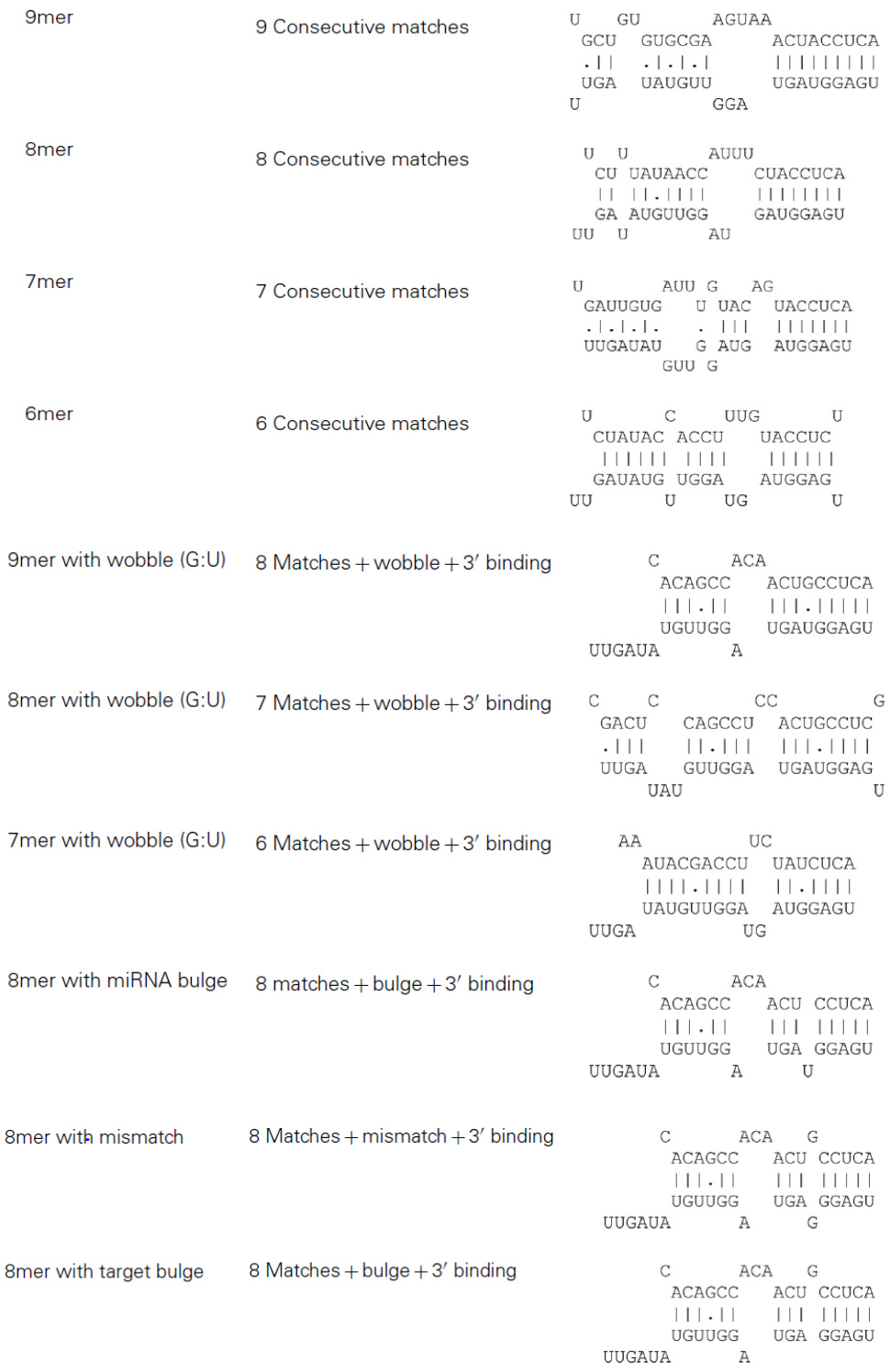
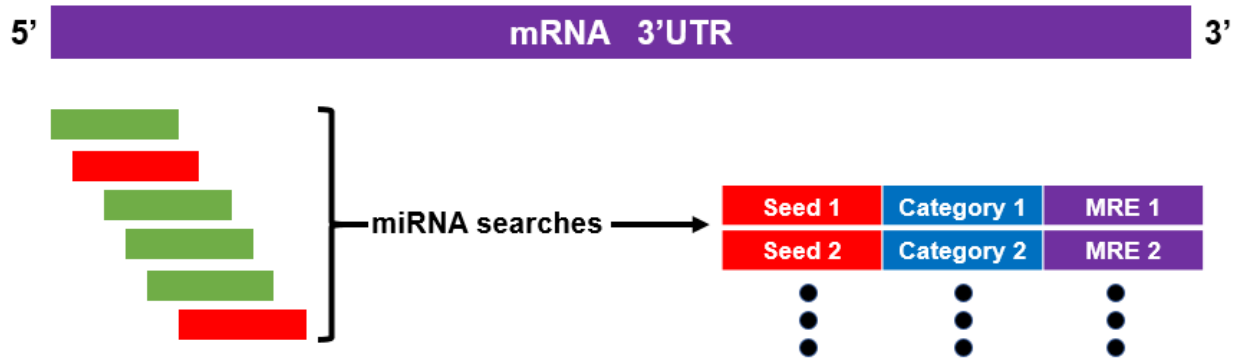


Figure 3. Binding categories for miRNA:mRNA interaction. Not included in this figure are 8mer with Mismatch, 8mer with Wobble, and target and miRNA bulges. These conditions were considered when we developed the algorithm to search for target sequences. (Reczko et al. 2012)




```
Imports
Define methods
Make dictionaries
Make matrix of miRNA seqs
Make matrix of mRNA seqs
For each mRNA seq:
      For each miRNA seq:
            Get complement
            Isolate extended seed
            Search for 9mer
            If match found:
                  Create profile and add to output file
            If bases 1-5 of MRE/extended seed match:
                  Search for 8mer
                  If match found:
                        Create profile and add to output file
                  Search for 7mer
                  If match found:
                        Create profile and add to output file
                  Search for 6mer
                  If match found:
                        Create profile and add to output file
```

Search for 8mer w/ Mismatch @ 5th position

If match found:

Create profile and add to output file

Search for 9mer w/ Wobble @ 5th position

If match found:

Create profile and add to output file

Search for 8mer w/ Wobble @ 5th position

If match found:

Create profile and add to output file

Search for 7mer w/ Wobble @ 5th position

If match found:

Create profile and add to output file

Search for 8mer w/ Mismatch @ 6th position

If match found:

Create profile and add to output file

Search for 9mer w/ Wobble @ 6th position

If match found:

Create profile and add to output file

Search for 8mer w/ Wobble @ 6th position

If match found:

Create profile and add to output file

Search for 7mer w/ Wobble @ 6th position

If match found:

Create profile and add to output file

If 1st base is mismatch and bases 2-6 match:

Search for 8mer

If match found:

Create profile and add to output file

Search for 7mer

If match found:

Create profile and add to output file

Search for 6mer

If match found:

Create profile and add to output file

Search for 8mer w/ Mismatch @ 5^{th} position

If match found:

Create profile and add to output file

Search for 9mer w/ Wobble @ 5^{th} position

If match found:

Create profile and add to output file

Search for 8mer w/ Wobble @ 5^{th} position

If match found:

Create profile and add to output file

Search for 7mer w/ Wobble @ 5^{th} position

If match found:

Create profile and add to output file

Search for 8mer w/ Mismatch @ 6^{th} position

If match found:

Create profile and add to output file

Search for 9mer w/ Wobble @ 6^{th} position

If match found:

Create profile and add to output file

Search for 8mer w/ Wobble @ 6^{th} position

If match found:

Create profile and add to output file

Search for 7mer w/ Wobble @ 6^{th} position

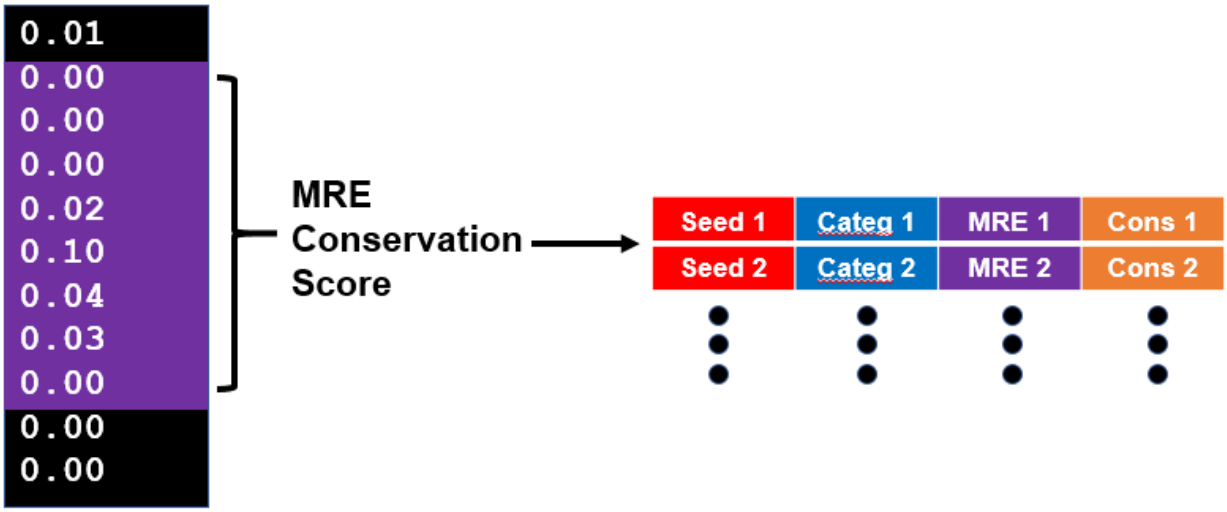
If match found:

Create profile and add to output file

Print category counts

EOF

Figure 4. Seed Sequence Search Algorithm.




```
Imports

For each profile:

      Use extended_seed and mre as input for RNAhybrid

      Add mfe and pval to profile

      Add updated profile to output file
```

Figure 5. Binding Energies Algorithm.

target	target_range	mirna	seed	categ	MRE	cons	mfe	pval
NR_030382	chr1:10015461 1-100178513	miR-7-1-3p	CAACAAUC	8merTB	ATTAGGTTG	7.224	-6.9	1.000
NR_030382	chr1:10015461 1-100178513	miR-7-2-3p	CAACAAUC	8merTB	ATTAGGTTG	7.224	-6.9	1.000
NR_030382	chr1:10015461 1-100178513	miR-140-3p	UAACACAGG	6mer	ATTGTGGTT	6.208	-13.9	0.001

```
Imports
Convert chr cons score files to list of lines
Make list of [chr,start_position,end_position] from list of lines
Make new file of cons scores with positions
For each line in new file:
      Open respective file and add scores to score_list
For each chr:
      Search score_list and add scores
For each chr:
      Add cons score to respective profile
      Add updated profile to output file
```

Figure 6. Conservation Score Algorithm.

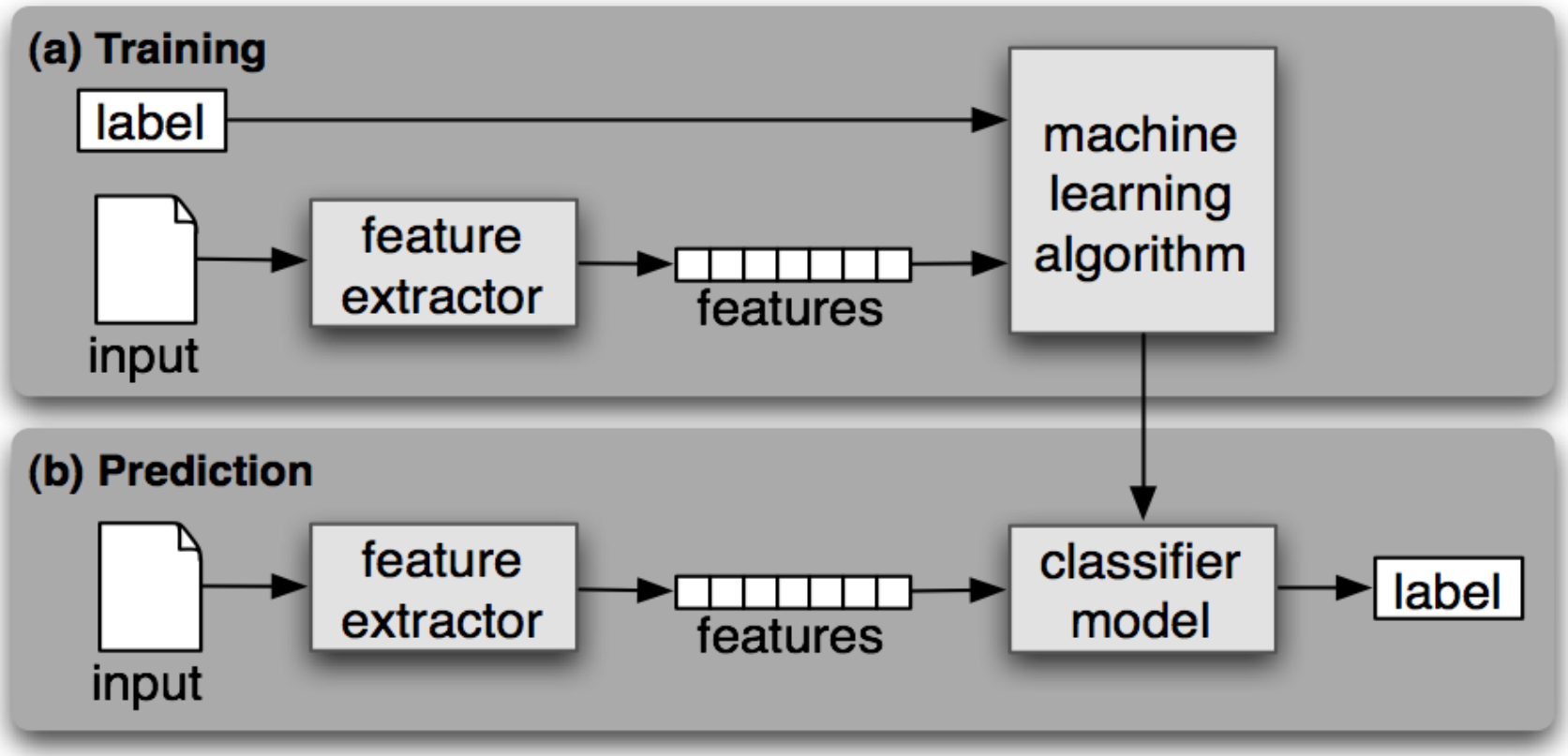


Figure 7. The development of a machine learning model for target prediction. Once profiles are made, training of the model will be done to check for its accuracy (a). When training is complete and satisfactory, prediction of unknown novel targets will proceed (b), after which verification with wet-lab experimentation will be needed.

Table 1. This table shows the number of profiles generated by our algorithm for each binding category. In all, there were 1,796,319 profiles generated.

Binding Category	Number of Profiles Generated
9mer	3,161
8mer	16,539
7mer	60,718
6mer	231,063
9mer w/ Wobble	39,184
8mer w/ Wobble	155,270
7mer w/ Wobble	633,079
8mer w/ Mismatch	657,305

References

Cambronne, X. A., Shen, R., Auer, P. L., & Goodman, R. H. (n.d.). Capturing microRNA targets using an RNA-induced silencing complex (RISC)-trap approach. https://doi.org/10.1073/pnas.1218887109

Cheng, S., Guo, M., Wang, C., Liu, X., Liu, Y., & Wu, X. (2015). MiRTDL: A deep learning approach for miRNA target prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2015.2510002

Chi, S. W., Zang, J. B., Mele, A., & Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature*. https://doi.org/10.1038/nature08170

Gomes, C. P. C., Cho, J. H., Hood, L., Franco, O. L., Pereira, R. W., & Wang, K. (2013). A review of computational tools in microRNA discovery. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2013.00081

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., … Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*. https://doi.org/10.1016/j.cell.2010.03.009

Krol.Filipowicz.2010_reg.miRNAbiogen.func.decay_nrg2843. (n.d.).

Krüger, J., & Rehmsmeier, M. (2006). RNAhybrid: MicroRNA target prediction easy, fast and flexible. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl243

Morita, S., Nakabayashi, K., Kawai, T., Hayashi, K., Horii, T., Kimura, M., … Hatada, I. (2016). Gene expression profiling of white adipose tissue reveals paternal transmission of proneness to obesity. *Nature Publishing Group*. https://doi.org/10.1038/srep21693

Mu, W., & Zhang, W. (2012). Bioinformatic resources of microRNA sequences, gene targets, and genetic variation. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2012.00031

Reczko, M., Maragkakis, M., Alexiou, P., Papadopoulos, G. L., & Hatzigeorgiou, A. G. (2012). Accurate microRNA target prediction using detailed binding site accessibility and machine learning on proteomics data. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2011.00103

Rehmsmeier, M., Steffen, P., Höchsmann, M., Chsmann, M. H., & Giegerich, R. (2012). Fast and effective prediction of microRNA/target duplexes BIOINFORMATICS Fast and effective prediction of microRNA/target duplexes. *Cold Spring Harbor Laboratory Press on March*, *1*. https://doi.org/10.1261/rna.5248604