Prediction of Target Binding of microRNA Molecules using Machine Learning Approaches

Daniel Stribling[*]
Dept. of Molecular Genetics & Microbiology; UF Genetics Institute
University of Florida
Gainesville, FL 32610
ds@ufl.edu

Shaunak Sompura
Dept. of Computer and Information Science and Engineering;
Herbert Wertheim College of Engineering
University of Florida
Gainesville, FL 32611
sompura.shaunak@ufl.edu

Abstract

1 Introduction

1.1 Central Dogma of Molecular Biology

In the past few decades, the field of genomics (the study of DNA genomes) has leaped forward utilizing powerful new computational techniques. The function of deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and the worker protein molecules are linked by the central dogma of molecular biology:

$$DNA \xrightarrow{makes} RNA \xrightarrow{makes} Protein$$

Figure 1: The "Central Dogma" of molecular biology. An organism's DNA genome encodes RNA (the transcriptome), which in turn encodes for protein (the proteome) which primarily perform the activities that sustain life.

DNA molecules are primarily comprised of a linear, ordered sequence of nucleotide "bases" from the set: $\{A, T, C, G\}$. These molecules are primarily found in long, double-stranded, complementary pairs, with each nucleotide found in opposition to a known partner: $\{A/T,\ G/C\}$ These long-lasting DNA molecules then serve as a template for the shorter, active RNA molecules that have the corresponding bases: $\{A, U, C, G\}$.[2]

1.2 microRNA

microRNA (miRNA) are RNA molecules that are much shorter than the protein-coding RNAs of the central dogma, measuring 18-22 nucleotides in length. They serve an important regulatory function

[*] http://www.github.com/dstrib ; https://www.linkedin.com/in/danielstribling/

[2] For example, the DNA sequence $ATCGTC$ acts as a template for the RNA molecule: $AUCGUC$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

in human cells by interrupting the process of the central dogma 1. The nucleotide sequence of a miRNA determines its targets. When a targeting miRNA finds a complementary protein-coding RNA, it triggers the blockage or destruction of that coding RNA. The exact specific sequence properties required for target-recognition are not known, and machine learning methods are being used for target prediction.

1.3 microRNA Raw Target Data

The structure of microRNA target prediction question is formulated as such. Each raw datapoint consists of:

- an approximately 22 nucleotide-length sequence (a miRNA)
- a nucleotide sequence of approximately 22 nucleotides, representing a subsequence of a potential target sequence

A simple goal for this prediction is to determine whether a microRNA will bind with sufficient strength to a prospective target to regulate the target's function. This binding is primarily, but not exclusively based on 1:1 positional sequence complementarity of matching bases. Letting n_i represent the nucleotide position index of the miRNA, and n_j represent the nucleotide position index of the target, a complementary (matching) sequence would be:

$$1_i : A \;/\; 1_j : T, \quad 2_i : C \;/\; 2_j : G, \quad ...$$

Figure 2: An example sequence with full complementarity

The trivial cases are complete sequence complementarity, which provides the strongest possible miRNA/target binding, and absent sequence complementarity which provides the weakest possible binding. The majority of both positive and negative pairs have intermediate complementarity, as statistically about 4-5 indices will be complementary based only on random chance. The goal of miRNA target identification is to predict the binding based on the respective miRNA and target sequences and the pattern of complementarity between those sequences.

1.4 Problem Formulation

The goal for miRNA prediction is to minimize the difference in a specific target parameter relating to a given miRNA target pair. This parameter can either be continuous, like "binding strength", or can be categorical, such as "bound or unbound". The "binding "strength" parameter can be assigned to be categorical via a cutoff threshold. The problem is formulated as follows:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\phi_i)$$

Figure 3: miRNA Optimization Formulation. Where n is the number of samples, θ is the function parameters, and ϕ_i is the values of the features for data point i

1.5 microRNA Target Prediction Features

To date, the most successful miRNA target identification algorithms have been based on extracting features based on the pattern of complementarity between two sequences, and rely little on the specific nucleotide identities within the sequences themselves. The prediction feature is complete (6-bases) or near-complete (5-bases) nucleotide complementarity in the "seed region" of the miRNA from indices 2 to 7, (1,2) although a recent study has shown that strong complementarity in higher indices can favor binding without a strong seed match (2).

A 2015 study utilized four machine learning algorithms and identified thirteen features to be useful in prediction. (3) Described here are 9 (of 13) features that depend directly on the two described sequences, and not on specific score calculations or external biological databases:
(i) energy of folding predicted by an external program, (ii) seed sequence match, (iv) AU nucleotide content, (viii) the total number of paired positions, (ix) the length of the target region, (x) the length of the largest consecutive pairings, (xi) the position of the largest consecutive pairings relative to the start of the miRNA; (xii) the number of paired positions at the miRNA end (within the last 7 indices), (xiii) the difference between the number of paired positions in the seed region and that in the miRNA end. Of note, only two of the 13 features: (i) folding energy, and (iv) AU nucleotide content, are based on the specific nucleotides in the miRNA sequence and the remainder are based on patterns of complementarity between sequences.

2 Algorithm

2.1 Least Squares Regression

Several algorithms have been used over the last decade to determine miRNA / target binding. The most widely used algorithm for prediction of miRNA / target binding is implemented as the TargetScan software by the Bartel group. (1) This method uses multiple regression of features to predict binding with candidate targets. The features are primarily based on seed sequence matches and patterns of biological relevance. (2)
In this case, the problem formulation is given with the loss function defined as least squares regression or linear regression with multiple variables:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\phi_i) = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \phi_i^\top \theta)^2$$

$$Let: \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \Phi = \begin{bmatrix} \phi_1^\top \\ \phi_2^\top \\ \vdots \\ \phi_n^\top \end{bmatrix}$$

$$Then: \quad \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \phi_i^\top \theta)^2 = \min_{\theta} \frac{1}{n} \|y - \Phi\theta\|_2^2$$

For the optimal $\hat{\theta}$, where the gradient is set to zero:

$$\Phi^\top \Phi \hat{\theta} = \Phi^\top y$$

Thus, given that $\Phi^\top \Phi$ is nonsingular, the optimal weights $\hat{\theta}$ can be found by:

$$\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top y$$

Figure 4: Least Squares Regression Formulation for miRNA target prediction. Where n is the number of samples, θ is the function parameters, and ϕ_i is the values of the features for miRNA / target pair i

2.2 Least Squares Regression with LASSO Regularization Algorithm

A 2015 study by Ding et al. implemented 4 methods to study the described features in a package called TarPmiR. (3) This method utilized a greater number of data features than TargetScan, and also added LASSO regularization to highlight the most significant features utilized for miRNA / target prediction. (3)

Here, the problem formulation is given with the loss function defined as least squares regression with an added term for regularization of the parameter weights:

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\phi_i) = \min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi_i^\top \theta)^2 + \lambda \sum_{j=1}^{n} |\theta_j|_1$$

$$Let: \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \Phi = \begin{bmatrix} \phi_1^\top \\ \phi_2^\top \\ \vdots \\ \phi_n^\top \end{bmatrix}$$

$$Then: \quad \min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi_i^\top \theta)^2 + \lambda \sum_{j=1}^{n} |\theta_j|_1 = \min_{\theta} \frac{1}{n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_1$$

The LASSO method does not have a closed-form solution, however a guaranteed solution exists as the LASSO function above is convex in paramter θ.

Figure 5: LASSO Least Squares Regression Formulation for miRNA target prediction. Where n is the number of samples, θ is the function parameters, ϕ_i is the values of the features for miRNA / target pair i, and λ is the regularization weight.

2.3 Support Vector Machine

The support vector machine (SVM) machine learning method was implemented by the Wang group, in which miRNA targeting features are modeled in a SVM framework. (4) A recursive feature elimination (RFE) analysis is performed to rank the relative importance of each feature for its independent contribution to model performance. In this RFE evaluation, all the features are analyzed collectively using SVM. The final SVM model utilizes 96 features to build a prediction model named MirTarget (4). The derivation of the SVM basis is shown in Figure 6.

$$\min_{\omega,\xi,b} \frac{1}{2}\omega^T\omega + C\sum_{k=1}^{m}\xi_k$$

subject to

$$yk(\omega^T\phi(x_k) + b) \geq 1 - \xi_k$$

$$\xi_k \geq 0, k = 1, ..., m.$$

$$f(x) = sign(\omega^T\phi(x) + b)$$

Figure 6: The SVM algorithm learns by training a hyperplane (ω, b) to separate two classes where x is the training vectors and y is the target vector. Classification is based on the SVM score calculated by $f(x)$

A different example of the application of the SVM method is MultiMiTar, an SVM based classifier integrated with a multi objective metaheuristic-based feature selection technique. The method uses high quality negative examples and selection of biologically relevant miRNA-targeting site context specific features with a novel feature selection technique AMOSA-SVM with multi objective-optimization (MOO). AMOSA-SVM extracts a set of informative, nonredundant features that enhance the predictive power of the proposed classifier MultiMiTar. The SVM is used as the classifier by transforming the input data into another higher dimensional feature space where it is easy to compute an accurate classification. (5)

3 Experiments

3.1 Dataset Preparation

For implementation of each of the tested algorithms, the dataset used for training of the TarPmiR algorithm (3) was selected. This dataset provides a set of positive examples of miRNA and targeted sequences published by the Helwak group that result from an experimental method that provides miRNA paired to known targets. (6) The data was downloaded from the GEO database, accession number GSE50452. The data was provided in the ".hyb" genomic sequence format. Data manipulation and implementation of sequence features was performed in Python3. The hybkit software package was used to perform processing of data in ".hyb" format. (7) The Vienna software package was used for the calculation of the free energy for each miRNA target pair.

As the dataset only provides examples of positive features, a set of negative sequence pairs was created for training. For each positive miRNA / target pair, a corresponding unmatched miRNA/target sequence was created. A random site on the same target molecule was chosen to replace the original positive target site. For all data points, Each of the 9 sequence based features listed above were calculated and output in csv format. In addition, 3 sequence based features based on the above features were also added. For a complete list of feature titles, see Table 1.

For both the least squares and LASSO regression methods, random-choice K-fold cross validation was performed with K = 10. Thus, the dataset was randomly partitioned into 10 randomly-selected training / test datasets with a 9:1 training:test ratio.

3.2 Least Squares Regression

As the least squares regression model has been utilized in the standard for miRNA target prediction: TargetScan (1), a least squares regression model was implemented over the feature dataset. Bound pairs were assigned a value of 1, and unbound pairs were assigned a value of -1. The loss function shown in Figure 7 was then used for training. The prediction for data θ_i was then found for each binding point determined as shown in Figure 8.

$$\ell(x_i) = \|y_i - x_i^\top \theta + \beta\|^2$$

Figure 7: Least Squares Loss Function

$$prediction = sign(x_i^\top \theta + \beta)$$

Figure 8: Least Squares Data Prediction

The average prediction accuracy achieved across all cross-validated training dataset partitions was 84.47%, a very high accuracy for the miRNA prediction field. The results for each individual run are shown in Figure 9.

3.3 LASSO Regression

Another method utilized in a popular model is the LASSO Regression method, implemented by the developers of TarPmiR (3). The LASSO regression model was implemented over the feature dataset. The loss function shown in Figure 10 was then used for training. The prediction for data θ_i was then found for each binding point determined as shown in Figure 11. Dataset preparation and partitioning were performed as with the Least Squares Regression method.

Several values of λ were tested during the implementation process, and a value of of $\lambda = 0.1$ was chosen to balance feature removal and accuracy. With this value, the prediction accuracy achieved

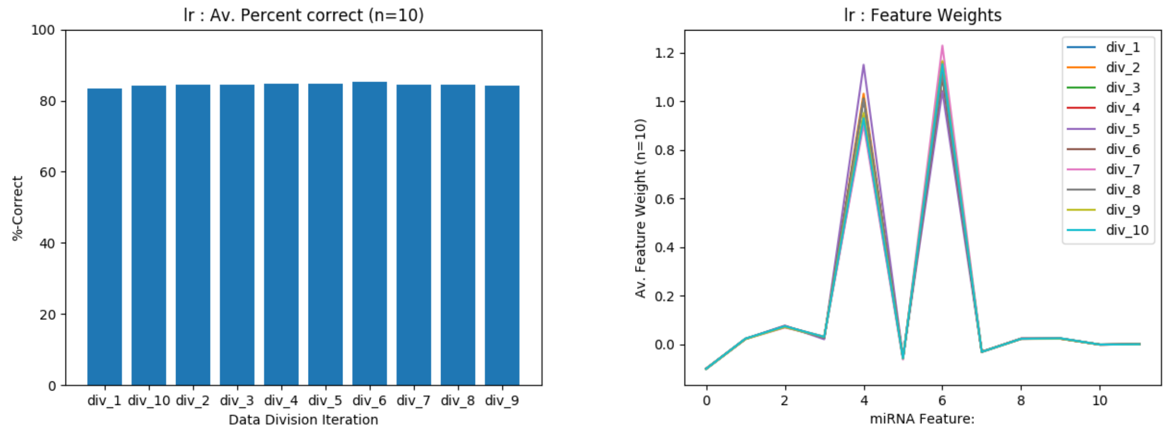


Figure 9: Results achieved for least squares regression over 10 partition iterations. Left: Prediction Accuracy. Right: Feature Weight

$$\ell(x_i) = \|y_i - x_i^\top \theta + \beta\|^2 + \lambda\|\theta\|_1$$

Figure 10: Least Squares "LASSO" Loss Function

$$prediction = sign(x_i^\top \theta + \beta)$$

Figure 11: Least Squares "LASSO" Data Prediction

was highly comparable to the accuracy achieved by least squares regression at 83.97% (compared to 84.47%). Notably, this result was achieved using only four features, as can be seen with the individual results shown in Figure 12 (right).

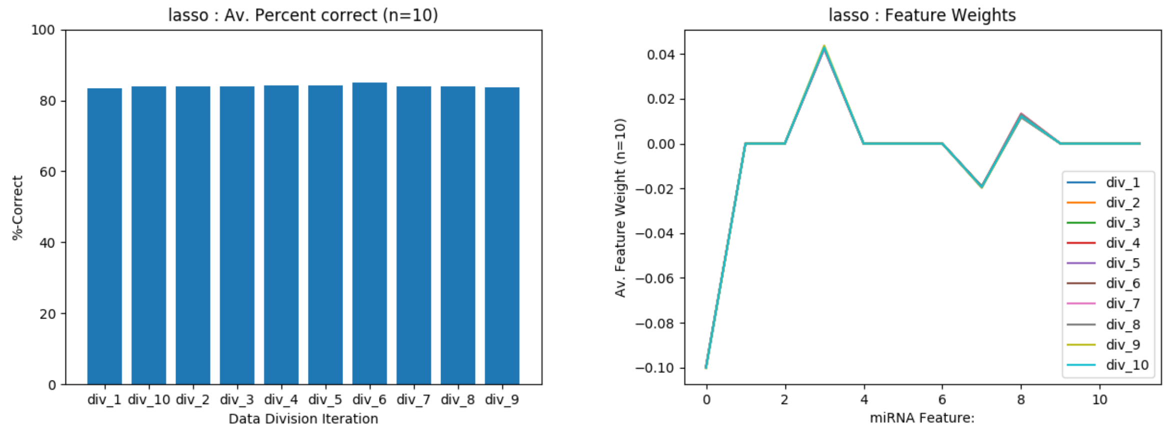


Figure 12: Results achieved for least squares regression over 10 partition iterations. Left: Prediction Accuracy. Right: Feature Weight

3.4 Support Vector Machine

The implementation of SVM algorithm yields an accuracy of 85.06% with a Gaussian Kernel whereas using a Linear Kernel has an accuracy of 84.50%. The more accurate RBF kernel has the ability to create non linear decision boundary based on high dimensional feature mapping. This is a major

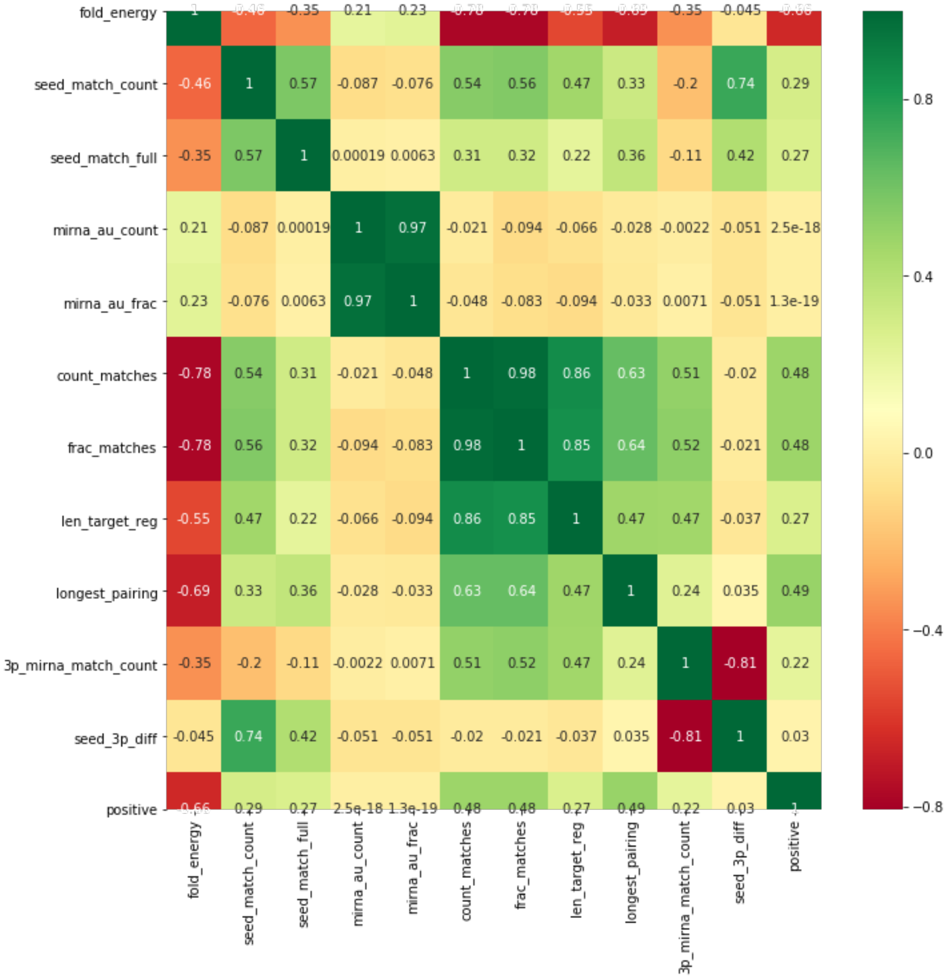
advantage over other linear boundary based approach when the is not clearly separable.
Feature selection techniques are utilized to understand the correlation amongst the feature vectors and the target vector as shown by the heatmap in Figure 15 along with Feature vector scores.

$$f(x) = sign(\omega^T \phi(x) + b)$$

Figure 13: SVM Decision Value f(x)

$$c(x, y, f(x)) = (1 - y * f(x))$$

Figure 14: SVM Loss Function



	fold_energy	seed_match_count	seed_match_full	mirna_au_count	mirna_au_frac	count_matches	frac_matches	len_target_reg	longest_pairing	3p_mirna_match_count	seed_3p_diff	positive
fold_energy	1	-0.46	-0.35	0.21	0.23	-0.78	-0.78	-0.55	-0.69	-0.35	-0.045	-0.66
seed_match_count	-0.46	1	0.57	-0.087	-0.076	0.54	0.56	0.47	0.33	-0.2	0.74	0.29
seed_match_full	-0.35	0.57	1	0.00019	0.0063	0.31	0.32	0.22	0.36	-0.11	0.42	0.27
mirna_au_count	0.21	-0.087	0.00019	1	0.97	-0.021	-0.094	-0.066	-0.028	-0.0022	-0.051	2.5e-18
mirna_au_frac	0.23	-0.076	0.0063	0.97	1	-0.048	-0.083	-0.094	-0.033	0.0071	-0.051	1.3e-19
count_matches	-0.78	0.54	0.31	-0.021	-0.048	1	0.98	0.86	0.63	0.51	-0.02	0.48
frac_matches	-0.78	0.56	0.32	-0.094	-0.083	0.98	1	0.85	0.64	0.52	-0.021	0.48
len_target_reg	-0.55	0.47	0.22	-0.066	-0.094	0.86	0.85	1	0.47	0.47	-0.037	0.27
longest_pairing	-0.69	0.33	0.36	-0.028	-0.033	0.63	0.64	0.47	1	0.24	0.035	0.49
3p_mirna_match_count	-0.35	-0.2	-0.11	-0.0022	0.0071	0.51	0.52	0.47	0.24	1	-0.81	0.22
seed_3p_diff	-0.045	0.74	0.42	-0.051	-0.051	-0.02	-0.021	-0.037	0.035	-0.81	1	0.03
positive	-0.66	0.29	0.27	2.5e-18	1.3e-19	0.48	0.48	0.27	0.49	0.22	0.03	1

Figure 15: A correlation heat map for the feature vectors is created to uncover the trends of vectors that affect the target vector the most.

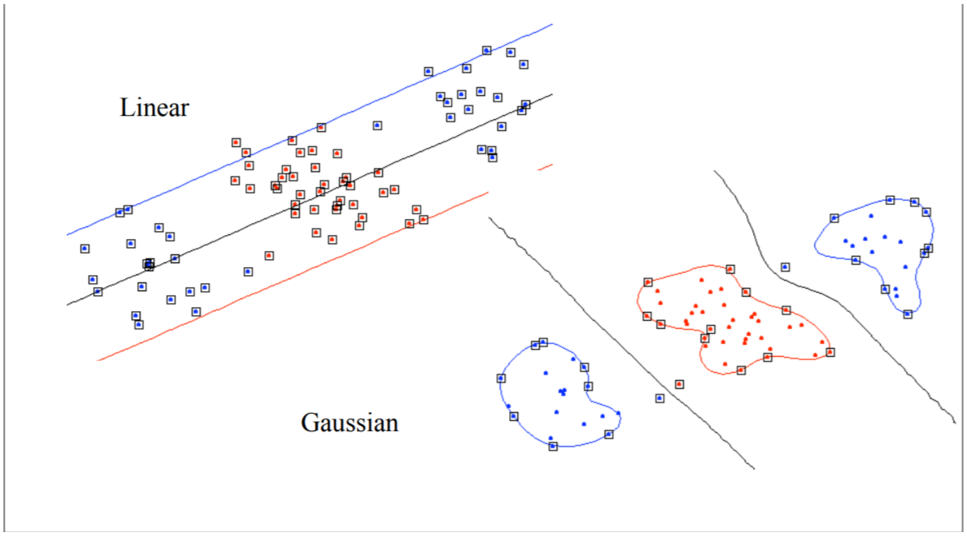


Figure 16: Comparision of Linear SVM model and Gaussian SVM model (9)

4 Conclusion

4.1 Interpretation of Prominent Features in Least Squares / LASSO

The least squares regression and LASSO implementations showed that a few features are primarily responsible for the prediction results observed, as many of the included features can be removed without a significant loss of accuracy. The average feature scores for the least squares and LASSO regressions are shown in Table 1.

parameter	least_squares	LASSO
fold_energy	-0.10129	-0.09993
seed_match_count	0.02204	0.0
seed_match_full	0.07344	0.0
mirna_au_count	0.02764	0.04271
mirna_au_frac	0.99317	0.0
count_matches	-0.05712	0.0
frac_matches	1.13188	0.0
len_target_reg	-0.03108	-0.01940
longest_pairing	0.02279	0.01242
3p_mirna_match_count	0.02344	0.0
seed_3p_diff	-0.00140	0.0
constant	0.0	0.0

Table 1: Least Squares and LASSO Average Feature Weights

Comparing these results to the analysis shown in the correlation heat map in Figure 16, it can be seen that the multiple variants of the same feature ("mirna_au_count" / "mirna_au_frac", and "count_matches" / "frac_matches") strongly correlate to each other. Thus it makes intuitive sense that only one of these correlated features will be utilized by the lasso prediction. Similarly the strong negative correlation between "3p_mirna_match_count" and "seed_3p_diff" indicates that only one of these features would be useful, and neither are selected by the lasso regression.

It can be seen that the features of greatest importance for prediction are "fold_energy," "mirna_au_count", "len_target_reg", and "longest_pairing." As the values for fold_energy are nonpositive, a negative weight corresponds to a positive contribution to folding. Interestingly, the "len_target_reg" feature contains only nonnegative values. This feature represents the difference between the index of the first bound position, and the last bound position: the "width" of the total area over which binding occurs. Thus, the negative weight indicates that the length of the total bound area contributes negatively to the probability that the target is bound. Using the positive contribution of the "longest_pairing" in both implementations, and "frac_matches" in the LASSO implementation, this suggests that short runs of continuous sequence matching contribute significantly to a positive pair more than simply the total number of matched bases.

4.2 Accuracy and Limitations

The accuracy achieved by all tested methods are >= 83% in distinguishing bound pairs from unbound pairs. Based on the ambiguity of a "true positive" dataset for the targets of human miRNA, there is no standard comparison by which to test the accuracy of the predictions. Different standard methods have proposed different validation tests. The TargetScan method directly parameters their validation in terms of experimentally determined results of miRNA repression activity. (1) Alternatively, the TarPmiR package parameterizes successful prediction in multiple ways including representation of target sequences in indirectly-correlated miRNA-target pairs, that were themselves identified to interact via a set of empirically generated rules. This method also determines accuracy using a large experimentally determined database of known miRNA / target pairs, but this database may not well represent binding patterns and as such comparison of prediction results to this database does not necessarily provide a ratio of accurate to inaccurate predictions. As such, it can be said that there is an 83% accuracy in predicting whether a provided pair is "True" or "False" using miRNA provided in the training dataset, and extension of this result to biological meaning is an ongoing challenge as discussed further below.

Opportunity for further prediction accuracy increase exists by manipulation of the negative dataset used for training. As negative data was generated artificially, an arbitrary amount can be created. In addition to the utilized 1:1 positive/negative ratio, additional 1:2, 1:4, 1:9, etc. ratios can be tested to determine the effect on prediction accuracy. Expansion from negatives resulting from "matched pairs" where there is a 1:1 correlation between a positive miRNA and a negative miRNA can also be changed to "truly random" negatives, in which a random miRNA is selected to interact with a random potential target. This would more truly represent the biological relevance, and would generalize the algorithm for prediction with all human miRNA as opposed to the 399 represented in the dataset.

4.3 Comparison to Other Methods

Recently, more advanced machine learning methods have been utilized and have been able to incrementally increase the accuracy of target prediction. As shown by the three successive implementations, using more advanced methods can provide some increase in accuracy. The TarPmiR group have additionally implemented a Random Forest approach to miRNA identification that has allowed up to 91% recall ($\frac{TP}{TP+FN}$) accuracy in target site identification. Thus, there is an opportunity to improve upon the $\sim 83\%$ accuracy achieved by the least squares and LASSO regression methods.

4.4 Generalizing to Biological Activity of miRNA

The true scientific question involved in miRNA targeting is the effect of a miRNA in biological systems. As such, the prediction of a miRNA must be characterized against the actual effect of that miRNA in cells. Recent experimental evidence by the Bartel lab (2) would provide a means to directly correlate the predictions made by the algorithm to miRNA effect, and thus characterize the predictions in terms of their potential biological relevance. Each prediction result would be correlated to the observed presence or absence of functional miRNA binding in cells. This could be further used to train the model, or used as a tool for validation of parameters used to design and select features or select an optimal training dataset.

References

[1] Agarwal, V., Bell, G.W., Nam, J., & Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs *eLife 2015, 4 e05005*

[2] McGeary, S.E., et. al. (2019) The biochemical basis of microRNA targeting efficacy *Science; 20 Dec 2019: Vol. 366, Issue 6472, eaav1741; DOI: 10.1126/science.aav1741*

[3] Ding, J., Li, X., & Hu, H. (2016) TarPmiR: a new approach for microRNA target site prediction *Bioinformatics 32.18 (2016): 2768-2775*

[4] Liu, W., & Wang, X. (2019) Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data *Genome Biology volume 20, Article number: 18 (2019)*

[5] Mitra, R., & Bandyopadhyay, S. (2011) MultiMiTar: A Novel Multi Objective Optimization based miRNA-Target Prediction Method *PLOS ONE 6(9): e24583. https://doi.org/10.1371/journal.pone.0024583*

[6] Helwak, A., & Tollervey, D. (2014) Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH) *Nature protocols 9.3 (2014): 711*

[7] Stribling, D., hybkit, (2020), GitHub repository, https://github.com/rennelab/hybkit

[8] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, CC., Stadler, P. F., and Hofacker, Ivo L., (2011) ViennaRNA Package 2.0 *Algorithms for Molecular Biology, 6:1 26, 2011, doi:10.1186/1748-7188-6-26*

[9] Apostolidis-Afentoulis, Vasileios. (2015). SVM Classification with Linear and RBF kernels. 10.13140/RG.2.1.3351.4083.