MiRNA Target Recognition and Expression Prediction From Customized Synthetic Datasets

Sumit Mukherjee[1], Randloph Lopez[2]

***Abstract*— This paper proposes a novel approach to micro RNA target recognition and prediction. Gene expression data is used for the same gene with different synthetically designed 3'UTRs. The possible targets are identified in the large data set of synthetic 3'UTRs for the commonly expressed micro RNA's for the particular cell line. The potential targets are then screened for structures which might prevent microRNA binding. A regression analysis is then performed to predict expression levels for different 3'UTRs. The problem is also studied from a different perspective. The synthetically altered portion of the 3'UTRs are screened for recurring k-mers. The k-mers corresponding to the lowest expressed mRNAs are identified and are compared with the seed sites of micro RNAs which show the most statistically significant effect on transcription.**

I. Introduction

The accurate prediction of microRNA seed target sites is essential for understanding their role in silencing gene expression. Significant efforts on this field have revealed that degenerate base-pairing is a common feature in RNA targeted binding [5]. Furthermore, it has been shown that reduced target accessibility due to secondary structure substantially reduces microRNA-mediated translational repression [2]. However, there is still a lack of understanding on the fundamentals that govern microRNA targeting. Existing research has been limited to study exclusively native microRNA targets and variations to the surrounding sequences of these binding sites. In contrast, we propose a novel method to explore the mechanism behind microRNA targeting by leveraging the use of randomly generated 20 base pair barcode sequences in the 3UTR of a reporter gene. Next-generation sequencing of the RNA reads corresponding to that gene reveals how the random sequence in the 3UTR affect gene expression after normalization to DNA reads of the plasmid to account for transfection variability. Further analysis reveals the specific contribution of particular k-mers in the barcode to microRNA gene silencing after association with microRNAs known to be present in the cell line population.

This project aims to identify and predict the level of micro RNA mediated repression in a particular cell line where the natively expressed microRNA's are known. A novel method of testing different sequences with barcoded 3'UTR for complimentarity with the different natively expressed micro RNA's is used here. Several different plasmids with the same Exon sequences and different intron and 3'UTR sites are used as seen in Figure 1. Each barcoded plasmid as transfected into HEK cells and their DNA and mRNA levels were measured after two days. Since transfection efficiencies would vary for each plasmid, the ratio of the mRNA to DNA counts was used as an indicator of expression level.

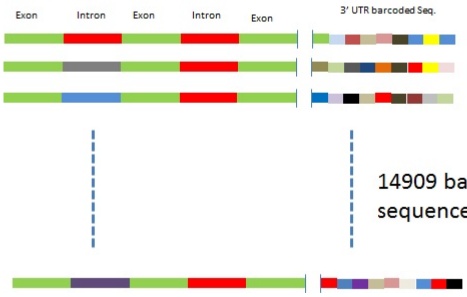


Fig. 1: Plasmids having the same Exon but with one different Intron each time and partly different 3'UTR sites are used

MicroRNA expression data for different cell lines was obtained from [1]. For the same cell lines micro RNA expression levels varied widely between different stages of cell growth. The average for the four different stages were calculated and only those micro RNA's with average count of 1 per cell were taken and the rest were rejected. There were 33 such microRNA's which met this requirement.

II. Results

A. Preliminary Target Site Recognition

The target site recognition was carried out by splitting each miRNA (converted to DNA) and barcoded sequence into k-mers. For the purposes of the analysis k = 7 was used because it is the average length of the micro RNA seed side. For each barcoded sequence, partial (1-mismatch was allowed) or complete complimentarity with the k-mers of each miRNA seed site was identified and recorded as a binary value is a matrix (A) i.e.

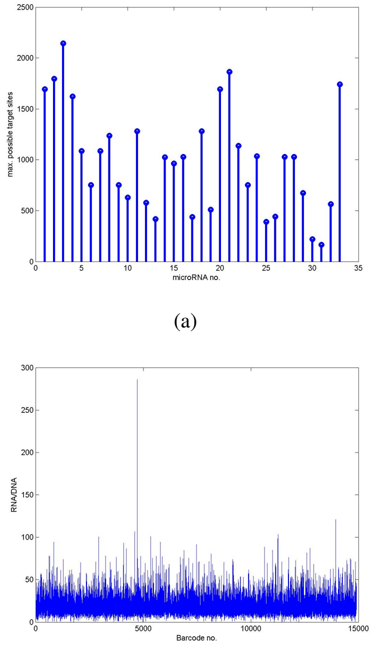
$$[A]_{i,j} = \begin{cases} 1 & \text{a k-mer of miR i = a k-mer of barcode j} \\ 0 & \text{otherwise} \end{cases}.$$

The number of target sites identified for each microRNA (both fully as well as partly complimentary) is shown in the Figure 2. The phenotype vector was simply the $\frac{mRNA_{count}}{DNA_{count}}$ for each barcoded sequence.

B. Statistical Significance of Data

The statistical significance of the data was tested using a QTL analysis. First the LOD score was calculated for each case as seen in Figure 3. Permutation test was then performed on each data set to obtain a distribution of LOD

[1] Department of Electrical Engineering, University of Washington, Seattle
[2] Department of Bioengineering, University of Washington, Seattle



(a)

(b)

Fig. 2: a) Stem plot for no. of targets for each miRNA b) Phenotype plot for each barcode

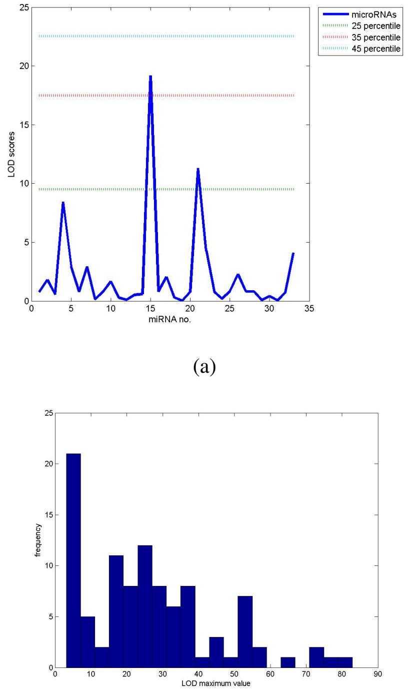
scores. For each case, various thresholds were used to obtain the statistically relevant micro RNA's. Owing to the fairly low quality of data, there were no statistically relevant microRNA's when the cut off percentile was kept too high. Only one microRNA (hsa-mir-92a-1) was seen to cross the 35 percentile mark.

C. *Structure Based Target Site Elimination*

It has been demonstrated in [2] that structures in the mRNA such as hairpins might prevent binding of the micro RNA to the target. This can be seen in Fig 4.

Mathematically, an equivalent statement is a potential target site is a viable one only if the Minimal Free Energy (MFE) structure of the hybrid has a lower energy than the MFE structure of the mRNA alone. That is $\Delta_s = MFE_{mRNA} - MFE_{hybrid} > 0$ for valid targets. However, in case of such structure, energy may also be required to unbind base pairs of the potential target sites which have to incorporated while considering the effect of structure. Hence, $\Delta\Delta_s = \Delta_s - \Delta_{unbinding} > 0$ for valid targets. Till now only the $\Delta_s > 0$ constraint has been implemented using the Vienna RNA package [3]. This yields more than a 20 percent reduction in number of targets as seen in Fig 5.

The QTL analysis on the new dataset yields a marked improvement in the LOD scores. This yields leads to two micro RNAs (hsa-mir-92a-1 and hsa-mir-185)with LOD scores over 35 percentile and approaching the 45 percentile mark as seen in Fig 6. The statistical significance of the dataset is expected to improve significantly once more spurious target sites are eliminated because of the energy of unbinding. However, this could not be implemented for the purpose of this class because of the time constraint.



(a)

(b)

Fig. 3: a) LOD score plot b) Distribution of maximum LOD score for randomized phenotype values

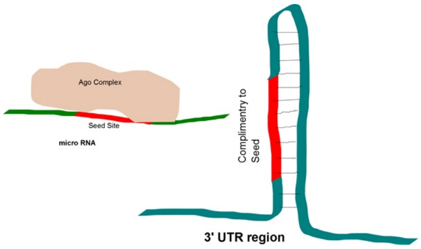


Fig. 4: Structures in mRNA 3'UTR region can prevent binding of micro RNA to target sites

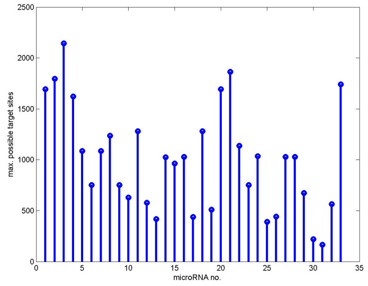
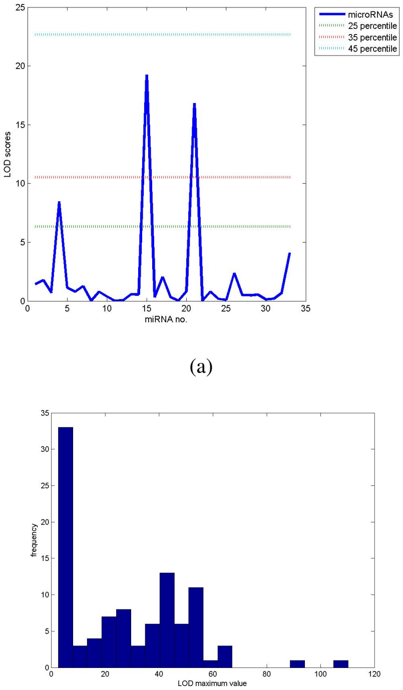


Fig. 5: Stem plot for no. of targets for each miRNA



(a)

(b)

Fig. 6: a) LOD score plot b) Distribution of maximum LOD score for randomized phenotype values

D. *Regression Analysis*

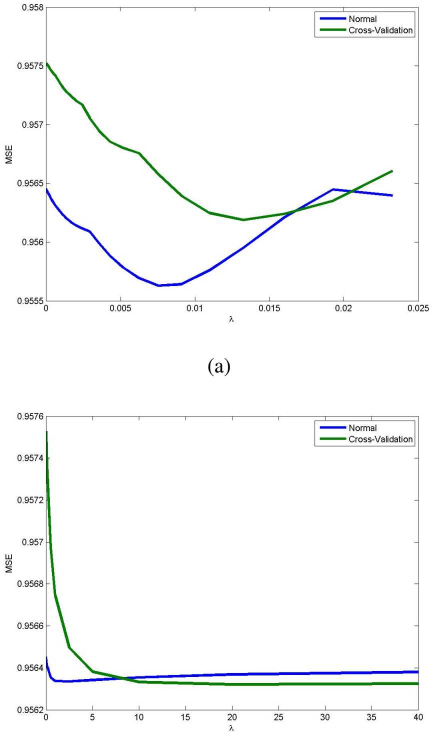
For an initial analysis it was assumed that the effect of repression of different microRNAs acts linearly upon the target i.e. their effects are additive. This can be expressed as :-

$$\hat{Y_i} = \sum_{j=1}^{j=N} a_j \hat{X_{ij}} \qquad (1)$$

Where, $\hat{Y_i}$ is the normalized expression level($\hat{Y_i} = \frac{Y_i - \frac{\sum_{i=1}^{i=N} Y_i}{N}}{\sigma_Y}$,$N$ is the number of micro RNAs, a_j is the weight for the jth micro RNA and $\hat{X_{ij}}$ is a similarly normalized value from the genotype matrix. Lasso and Ridge regression analysis were performed on the two data sets respectively. The resultant plots are shown in Figures 7 and 8. However it is seen that the MSE values in both cases are quite high even though there is a small improvement seen after considering the effects of structure. Since the statistical significance of the dataset was not very high, the bilinear and quadratic regression analysis were not performed. These will be performed at a later stage after incorporating the effect of the unbinding energy.

E. *Analysis of High Frequency K-mers*

An additional analysis was performed on the dataset without screening for the microRNAs known to be present in the cell line. This blinded analysis had the purpose of identifying 7-mers that had consistently lower expression levels than the rest and establishing the statistical significance of that difference. Initially, the 7-mers were extracted from the barcodes and those with a barcode repeat count of 22 or more corresponding to the 95th percentile- were including in the analysis. The histogram is seen in Fig 9.



(a)

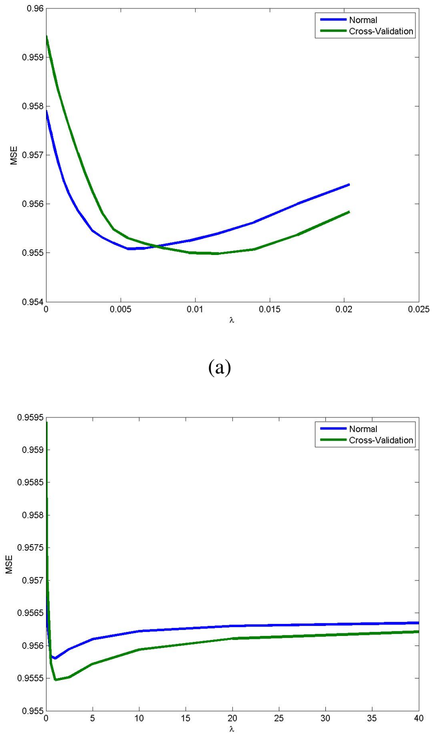
(b)

Fig. 7: Regression analysis plots for Target sites identified without considering structural effects: a) Lasso b) Ridge

The resulting 382 7-mers were subject to independent one sided t-test comparisons in order to establish whether lower repression levels were associated with any of them. Statistical significance was calculated by permutating the data set 100 times and calculating the lowest t-value. A t-value corresponding to the 95th percentile of this distribution of extreme t values (-14.77) was then used to establish statistical significance of the t-test results. This is seen in Fig 10.

T-test results indicate that none of the 7-mers evaluated showed statistically significant lower levels of RNA expression. We speculate that these results may be a consequence of a low size effect to biological noise ratio. This is evident from the broad distribution of t-values across the entire dataset which can be assumed to be product of the inherent variability introduced by RNA high throughput sequencing methods.Furthermore, the expression data in this analysis is only a function of microRNA mediated degradation of messenger RNA and it does not capture translation inhibition. Therefore, it is possible that by including an output measurement on translation inhibition we would be able to capture more significant repression levels.

The top 10 percentile of the k-mers (which correspond



(a)

(b)

Fig. 8: Regression analysis plots for Target sites identified with considering structural effects: a) Lasso b) Ridge

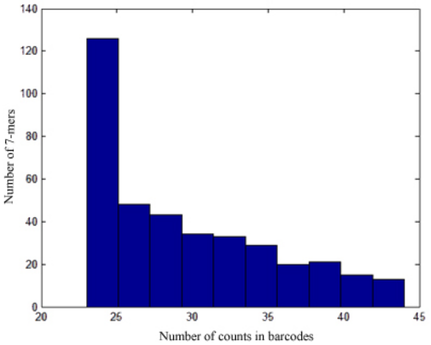


Fig. 9: Histogram of number of 7-mers present in multiple barcodes

to the lowest $\frac{mRNA}{DNA}$ ratio) are compared with the k-mers of all the micro RNAs expressed in the HEK cell line. It is observed that this has one of the k-mers matches with the seed site of the micro RNA hsa-mir-185, which is one of two micro RNAs whose LOD score surpasses the 35 percentile mark. This could indicate that there is some statistical significance in the dataset.

F. Validation on Test Case

To test that the target identification algorithm works the micro RNA hsa-mir-15a was chosen and three known 3'UTRs were chosen which are known to have its target sites from [4]. The genes chosen are from 'fibroblast growth factor 2', 'myotubularin related protein 3' and 'katanin p60 subunit A-like 1'. The first two are the highest scoring target sites from the list of the known target sites with a target score of 100 while the third has a mid level target with a target score of 50. All three targets were identified using our method.

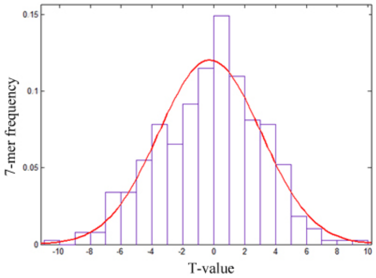


Fig. 10: Frequency of t-values corresponding to each individual 7-mer

III. Conclusion and Future Work

The low statistical significance of the dataset prevents good prediction of the expression levels. The cause of the low statistical significance could be narrowed down to biological noise and false target sites identified which can be correcting by eliminating more sites due to the effect of unbinding energy. The biological noise is harder to remove and possibly arises because of the difference in no. of transfected plasmids in each cell. Moreover, since the randomized barcode is only 20 nucleotides long and the rest of the 3'UTR is the same for all the plasmids, it can be assumed that the effect of repression is relatively small and hence gives rise to a small signal to noise ratio. This cannot be addressed with the present dataset since the experiments were not designed for the purpose of this analysis. In the future, entire 3'UTRs can be designed apriori to have multiple target sites of different natively expressed micro RNAs with checks for structural effects. By doing so, it is expected that we shall see larger effects of repression at the mRNA level and it would enable a better prediction of the expression level for any given 3'UTR. Moreover, in the future this approach can be used to study the effect of multiple target sites in the same 3'UTR. It has been suggested in [5] that the effect of multiple target sites of a micro RNA on the same 3'UTR may not be additive. Thus, it would necessitate the use of quadratic or bilinear or more complex models to capture their combined effects. Finally, this study provides a proof of principle for a new approach to predict expression levels of different genes.

IV. Acknowledgement

The authors are grateful to Alex Rosenberg and Dr. Georg Seelig of the Electrical Engineering Department of University of Washington for providing the dataset used in the study.

REFERENCES

[1] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, Amanda Rice, Alice O Kamphorst, Markus Landthaler, et al. A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414, 2007.

[2] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microrna target recognition. *Nature genetics*, 39(10):1278–1284, 2007.

[3] Vienna rna package. `http://www.tbi.univie.ac.at/RNA/`.

[4] mirdb micro rna target database. `http://mirdb.org`.

[5] Jennifer A Broderick, William E Salomon, Sean P Ryder, Neil Aronin, and Phillip D Zamore. Argonaute protein identity and pairing geometry determine cooperativity in mammalian rna silencing. *RNA*, 17(10):1858–1869, 2011.