Analysis of Paired miRNA-mRNA Microarray Expression Data Using a Stepwise Multiple Linear Regression Model

Yiqian Zhou[1], Rehman Qureshi[2], and Ahmet Sacan[3(✉)]

[1] Pure Storage, 650 Castro Street, Suite #260, Mountain View, CA 94041, USA
[2] Bioinformatics Facility, Wistar Institute,
3601 Spruce Street, Philadelphia, PA 19104, USA
[3] Biomedical Engineering, Drexel University,
3120 Market Street, Philadelphia, PA 19104, USA
ahmet.sacan@drexel.edu

Abstract. MicroRNAs are small endogenous RNAs that play important roles in gene regulation. With the accumulation of expression data, numerous approaches have been proposed to infer miRNA-mRNA regulation from paired miRNA-mRNA expression profiles. These mainly focus on discovering and validating the structure of regulatory networks, but do not address the prediction and simulation tasks. Furthermore, functional annotation of miRNAs relies on miRNA target prediction, which is problematic since miRNA-gene interactions are highly tissue-specific. Thus a different approach to functional annotation of miRNA-mRNA regulation that can generate context-specific expression levels is needed. In this study, we analyzed paired miRNA-mRNA expressions from breast cancer studies. The expression of mRNAs is modeled as a multiple linear function of the expression of miRNAs and the parameters are estimated using stepwise multiple linear regression (SMLR). We demonstrate that the SMLR model can predict mRNA expression patterns from miRNA expressions alone and that the predicted gene expression levels preserve differentially regulated gene sets, as well as the functional categories of these genes. We show that our quantitative approach can determine affected biological activities better than the traditional target-prediction based methods.

Keywords: Micro-RNA · Gene expression · Co-expression · Stepwise multiple linear regression

1 Introduction

MicroRNAs (miRNAs) are small (∼22 nucleotides) non-coding endogenous RNAs that play important roles in gene regulation by targeting the messenger RNA (mRNA) of protein-coding genes [1]. In most cases, though not always [2], miRNAs act to repress the expression of their target gene [3, 4]. miRNAs guide the repression by either degrading the mRNA molecules, decreasing the translational efficiency, or both. When a miRNA and its target mRNA are highly complementary, the pairing is extensive and the miRNA directs the cleavage of the mRNA, which is the predominant mode of

© Springer International Publishing AG 2017
Z. Cai et al. (Eds.): ISBRA 2017, LNBI 10330, pp. 59–70, 2017.
DOI: 10.1007/978-3-319-59575-7_6

miRNA-guided repression in plants. In animals, extensive miRNA-mRNA complementary pairing and the consequent cleavage of mRNA is less prevalent. Nevertheless, recent studies indicate that target mRNA degradation provides a major contribution to translational repression in animals [5, 6].

miRNAs participate in a wide range of biological processes, affecting the expression of over 60% of mammalian genes [7]. Over the past decade, it has become clear that miRNAs contribute to almost all known physiological and pathological processes, cancer being of particular interest. Since dysregulation of miRNAs is closely linked with dysregulation of oncogenes and tumor suppressors, studying the biological processes of miRNAs provides unique opportunities for the development of miRNA-based diagnostics and treatment of cancer [8, 9].

To understand the functions of miRNAs, a central goal and major challenge is to determine their target mRNAs. There are many experimental techniques for target identification of miRNAs of interest [10]. These experimentally identified miRNA-mRNA interactions are collected in several repositories, such as TarBase [11] and miRTarBase [12]. So far thousands of miRNAs have been identified in animals and plants, but only a small fraction of targets for these miRNAs have been validated experimentally, because of the low efficiency and high cost of experimental validation. Sequence-based computational methods have been developed to fill this gap by generating putative lists of miRNA-mRNA pairs, which have greatly reduced the number of interactions researchers need to validate experimentally. Widely used miRNA target prediction methods include TargetScan [7], miRanda [13], PicTar [14], TargetScanS [15], and DIANA-microT [16].

Currently, reliable prediction of miRNA-mRNA interactions remains a challenge. Predictions based solely on sequence information have high false positive rates [17]. In order to improve the performance, novel integrative approaches that combine sequence based predictions and miRNA experimental data are needed. Genome-wide mRNA expression measurement has become an indispensable tool in molecular biology. Similarly, technological advances have spawned a multitude of miRNA profiling platforms [18]. They together provide paired miRNA-mRNA expression profiles that enable researchers to pinpoint important miRNAs and their roles in particular biological processes.

Several methods that incorporate these high throughput data have been developed to find miRNA-mRNA regulatory pairs, including those based on correlation [19–22] or mutual information [23]. The findings from gene-expression analysis can be integrated with those from sequence-based methods by intersection [24] or weighted sum [20]. These simple approaches are efficient in extracting potential interactions from big datasets but they only consider independent pairwise miRNA-mRNA associations. Since a mRNA can be targeted by several miRNAs and its expression profile is affected by multiple miRNAs at the same time, multiple linear regression models have been proposed [25, 26]. When the data is co-linear or the number of samples is less than the number of regulators, the linear model is underdetermined and optimal solution is unattainable. This can be circumvented by introducing penalty terms to the system, such as $L_1 - norm$, $L_2 - norm$, or combination of both, of the coefficients of regulators [27]. In addition to regression-based approaches, several Bayesian models have been developed, inferring the posterior probability of real miRNA-mRNA interactions based

on the expression data, such as implemented in GenmiR++ [28] and its variations [29–31]. Bayesian network structure learning has also been proposed [32], in which regulatory relationships are represented as a graph and the graph that is best supported by the expression data is sought after.

The approaches proposed so far have focused on inference and validation of the "structure" of the miRNA-mRNA regulatory networks from the paired miRNA-mRNA expression data. Although knowing which genes are targeted by which miRNAs is of great value, it is not sufficient for determining whether a gene would be differentially expressed in a particular cellular context.

We have previously shown that a simple linear model is able to quantitatively predict and simulate gene expression levels in time-series data [33]. In this study, we investigate the application of a similar linear model for quantitative estimation of mRNA expression levels from miRNA data. The present study is unique in its focus on explicit quantitative modeling of gene expression levels, rather than just identifying miRNA targets.

2 Methods

We infer miRNA-mRNA regulatory interactions by analyzing paired miRNA-mRNA expression data using stepwise multiple linear regression (SMLR) [33]. Suppose there are M mRNAs and N miRNAs of interest; the expression level of each mRNA is modeled as a linear function of the expression levels of the miRNAs:

$$y_i = \beta_{i0} + \sum_{j=1}^{N} \beta_{ij} x_j + \varepsilon_i \tag{1}$$

where y_i and x_j are variables representing the expression of mRNA i and miRNA j respectively ($i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$); ε_i is the error term; and β_{i0} is a constant term representing the baseline mRNA expression. The β_{ij} term characterizes the regulatory effect of miRNA j on mRNA i. We identify the coefficient weights β_{ij} using stepwise multiple linear regression with a forward selection strategy, as described in our previous study [33]. Briefly, the predictors for a given gene y_i are identified starting with the inclusion of the constant term. In each forward selection step, individual predictor variables are considered for addition based on their statistical significance in the regression fitting. The p-value of an F-statistic for each variable is calculated to determine whether to include or exclude that variable in the model, using the null hypothesis that its weight coefficient is zero.

Suppose there are L samples; we can denote the expression of mRNA i and miRNA j across samples as row vectors: $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iL}]$ and $\boldsymbol{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jL}]$. More compactly, let $\boldsymbol{X} = [\boldsymbol{1}; \boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_N]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1; \boldsymbol{y}_2; \ldots; \boldsymbol{y}_M]$, with each row representing a mRNA or miRNA and each column representing a sample. If the data is already normalized, the constant term $\boldsymbol{1}$ in $\boldsymbol{X}$ can be dropped, leaving $\boldsymbol{X}$ and $\boldsymbol{Y}$ with dimensions of M-by-L and N-by-L, respectively, and representing the experimental data of M miRNAs and N mRNAs across L samples. Let $\boldsymbol{\beta_i} = [\beta_{i0}, \beta_{i1}, \beta_{i2}, \ldots, \beta_{iN}]$ and $\boldsymbol{B} = [\boldsymbol{\beta}_1; \boldsymbol{\beta}_2; \boldsymbol{\beta}_3, \ldots, \boldsymbol{\beta}_M]$. Then the SMLR model can be written in a simple matrix form:

$$\boldsymbol{Y} = \boldsymbol{B} * \boldsymbol{X} \tag{2}$$

The coefficient matrix $\boldsymbol{B}$ is M-by-N, which represents miRNA-mRNA regulatory interactions from M miRNAs and N mRNAs. Note that the coefficient matrix $\boldsymbol{B}$ is sparse, since the coefficients of insignificant interactions are set to zero.

Before estimating the interaction coefficients from training data and predicting gene expression levels, we need to perform necessary data pre-processing. Since we want to have a general model that works for expression datasets from different platforms and given the fact that most expression data available on Gene Expression Omnibus (GEO) database have already been normalized based on different assumptions regarding the specific platform, we avoid extra normalization across each sample unless necessary. First, we remove probes (genes) that have more than 3 missing data points and impute the missing value of the rest using the k-nearest-neighbor method with k = 3. Next, we center and scale the expression of each probe (gene) to have a mean value of zero and a standard deviation of one. This transformation does not alter the correlation between genes or the results of *t*-test for samples from different subgroups. Data preprocessing ensures that expression levels from different samples are on the same scale and that our predicted values can be directly compared with those from the real data. After preprocessing, we estimate the interaction coefficients $\boldsymbol{B}$ using stepwise multiple linear regression [33].

We evaluate the accuracy of the model predictions on both the training and independent testing datasets. In particular, we focus on how well the predictions preserve the differential expression profiles, as the list of differentially expressed genes is one of the most important outcomes from microarray studies. For both the real and predicted data, we perform Student's *t*-test to identify the genes that are significantly differentially expressed between experimental groups and analyze the overlap between the lists of genes generated from the real and predicted data.

A common downstream task in differential expression studies is the enrichment of differentially expressed genes into functional categories [34]. Here, we propose to use the mRNA levels estimated from our SMLR model for downstream functional annotation tasks. Considering any negative coefficient in the matrix $\boldsymbol{B}$ to indicate a targeting interaction, we evaluate the ability of our approach to discover mRNA targets and compare its performance to the TargetScan target prediction method [15] and to a negative correlation method where negatively correlated miRNA-mRNA are assumed to be targeting interactions (Pearson $p < 0.01$). Note that our method does not distinguish direct interactions from transitive ones or from those arising from co-regulation. Regardless of the source of the coefficients, our approach generates estimates of mRNA expression values, just as if they were obtained from a microarray gene expression experiment study. Once we obtain these estimated gene expression levels, we calculate a predicted list of differentially expressed genes and then perform gene set enrichment analysis using the DAVID web service [35]. Functional annotation is performed against OMIM, GO terms, BBID pathway, and KEGG pathway databases. We evaluate the performance on the functional enrichment task by comparing the resulting functional categories with those obtained from the real mRNA data and those obtained using target prediction methods.

In the following section, we first illustrate the application of SMLR to predict gene expression levels and functional categories, using a breast cancer expression profiling dataset. We then evaluate the ability of the model coefficients estimated from one dataset to generalize to another dataset generated from different experimental platforms. We compare the gene lists and functional categories predicted from miRNA data to those obtained from the real data and from TargetScan.

3 Results

In order to evaluate the ability of the SMLR model to predict gene expression levels from miRNA data, we first used the dataset available from a paired miRNA-mRNA study [36, 37], in which miRNA and mRNA profiles were obtained from the same primary breast cancer carcinomas (GSE19536, GSE22220), where the TP53 mutational and estrogen receptor (ER) status of each sample are also available. These samples are part of a larger cohort from the Oslo region [38].

After preprocessing, we obtained normalized expression profiles for 489 miRNAs and 40996 genes. We then performed leave-one-out-cross-validation (LOOCV) to evaluate the model, where we set aside one of the samples as the test sample and calculated the interaction coefficients from the remaining 100 training samples. The resulting model is then applied to the miRNA profiles from the training samples and the test sample separately. This procedure is repeated with each sample in the dataset used as the test sample.

Hierarchical clustering of the 1000 most differentially expressed mRNAs in the real data is shown in Fig. 1 (left). For comparison, a heatmap of the predicted expression levels are shown side-by-side (Fig. 1, right) with the same row and column arrangements. The predicted data displays surprisingly similar expression patterns, supporting

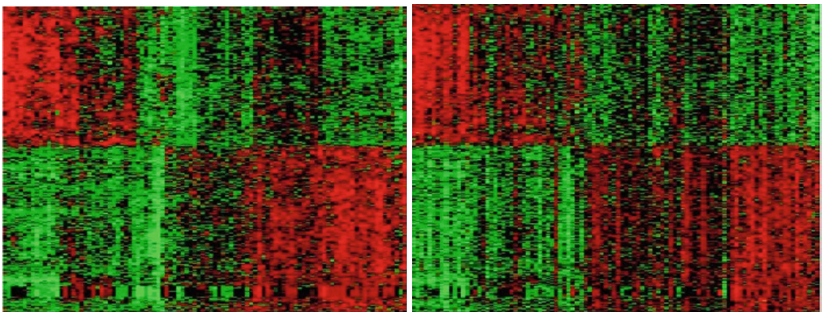

Fig. 1. Hierarchical clustering of mRNA expression. Left: Hierarchical clustering of the 1000 most differentially expressed mRNAs from the GSE19536 dataset. Right: expression levels of the same mRNAs predicted from the paired miRNA expression data, using SMLR with leave-one-out-cross-validation strategy. Rows are mRNA probes and columns are samples. Predicted data is shown with the same row and column arrangement as the real data. Root mean squared error (RMSE) of all predicted values was 1.11.

the idea that the miRNA expression alone provides a good summary of the gene expression state of the cell.

In order to further evaluate the reliability and usefulness of the gene expressions predicted from miRNA data, we examined whether the predicted values can identify a similar set of differentially expressed mRNAs. A two-sampled *t*-test on predicted gene expression data was performed between the ER-positive and ER-negative subgroups of samples. The p-values of the t-test are compared to those obtained from the original gene expression data (See Fig. 2-Left). These two set of p-values are highly correlated (r = 0.77). The mRNAs that are differentially expressed in the real data were likely to be found differentially expressed in the predicted data as well.

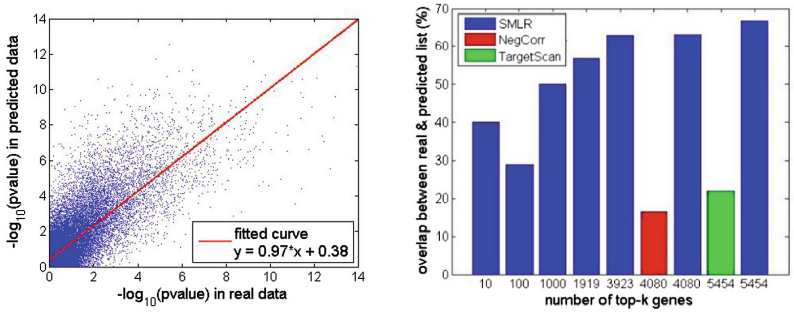

Fig. 2. Left: Comparison of differentially expressed mRNAs identified from the real and predicted expression data. Each point represents a mRNA, where the x and y axes show the $-\log_{10}$ transformed p-values obtained from an unpaired *t*-test in real and predicted data, respectively, comparing ER-positive and ER-negative breast cancer samples. The least-square fitted line is shown in red. **Right: Amount of overlap between the lists of differentially expressed genes in real and predicted data.** Percentage overlap between the most differentially expressed gene sets obtained from real and predicted data is shown. Each bar shows gene sets obtained with either a top-k or p-value criteria. After false discovery rate (FDR) correction, there were 1923 and 3942 mRNAs with p-value <0.01 and 0.05, respectively. (Color figure online)

Genome-wide microarray analysis is often used to prioritize a set of genes for follow-up wet-lab experimentation; such as reporter assays to confirm transcription, measurement of protein levels by northern blots, or knock-out experiments to evaluate phenotypic outcomes resulting from the absence of a gene. As such, it is important that our predictions preserve the ranking of the differentially expressed genes. Figure 2-Right shows the overlap between the top-k most differentially expressed gene sets obtained from the real and predicted data. The figure also shows the amount of overlap for gene sets obtained with the commonly used p-value thresholds of 0.01 and 0.05. At different top-k or p-value cut-offs, about half of the genes from the predicted gene set are in common with the real gene set.

Considering the noisy nature of gene expression data and the biological complexity of the rules governing translation of mRNAs to different protein isoforms, differential expression detected in microarray experiments is not conclusive for similar expression of the encoded proteins or for regulation of a particular phenotype the genes are

involved in. Gene set enrichment is commonly utilized to find biological functions affected by the concerted changes in a set of genes.

For a miRNA study, the functional annotations of miRNAs of interest can be obtained by enrichment analysis with a set of their target mRNAs. Traditionally, the set of miRNAs of interest are selected according to their differential expression patterns and their targets are selected from sequence-based target prediction algorithms or from experimentally validated targets. All targets of differentially expressed miRNAs are then (falsely) assumed to also be differentially regulated, even though these target genes are also targeted by other non-differentially expressed miRNAs. This is an unrealistic assumption that results in thousands of genes, limiting the statistical power of the enrichment analysis. This is demonstrated in Fig. 2-Right, where we compare the accuracy of the genes assumed to be differentially regulated from negative correlation and TargetScan predictions (17% and 22%, respectively) with those obtained from our method (63% and 67% for the same number of genes). Compared to context-agnostic target-prediction methods, we more effectively utilize the cellular context available from the state of all miRNAs in determining whether a gene is differentially expressed.

We performed functional annotation of the gene lists using DAVID [35]. For real data, which is used as the ground truth, and for SMLR, we used differentially expressed genes ($p < 0.01$) in real and predicted expression data, respectively. For other methods, the gene lists were formed by combining all of the targets of differentially expressed miRNAs ($p < 0.01$). Overlap of the functional annotation terms obtained from different methods with those generated from the real data are shown in Fig. 3-Right. Top-3 functional categories enriched from the real data were: Phosphoprotein, Alternative Splicing, and Splice Variant. SMLR was able to generate the same three terms in its top-3; whereas TargetScan and negative correlation only ranked only one of them in

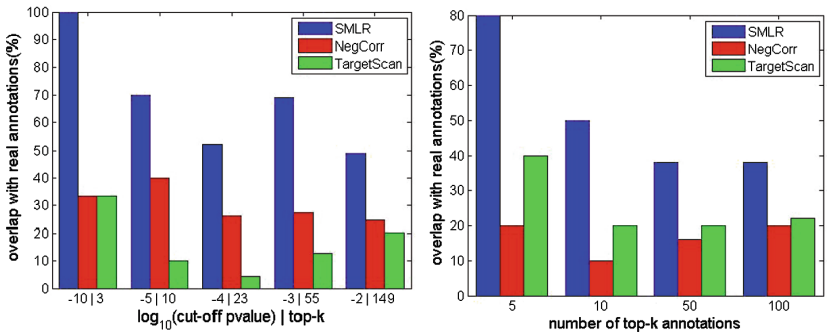

Fig. 3. Left: Functional enrichment from different methods. Percent overlap of functional annotations obtained from different methods with those obtained from real data are shown. At each p-value cutoff from SMLR, the same number of top-k annotations from each method are compared. Full list of enriched terms is available in the supplementary data. **Right: Comparison of functional enrichment in GSE19536 dataset.** SMLR is trained using GSE22220 dataset and differentially expressed genes from the predicted GSE19536 data are used for gene set enrichment. Negative correlation and TargetScan methods use all the predicted targets of differentially expressed miRNAs in GSE19536.

their top-3 lists. For the top-10 functional annotations obtained from each method, 70% were in common between results from real data and SMLR prediction, sharing similar rankings in statistical significance; while 40% and 10% were in common for negative correlation and TargetScan methods. These results support the claim that gene expression values predicted from miRNAs alone can capture the affected biological processes and that the functional annotations from estimated mRNA values are more accurate than those from collection of predicted targets.

The results above were obtained by leave-one-out cross-validation within a single experimental study, where each miRNA to mRNA mapping in a test sample was done using a model trained on the rest of the samples. Here we also evaluate the cross-database performance of SMLR by applying the model trained from one study to a dataset from an independent experimental study. Specifically, we train a model on GSE22220 dataset [36] and test its prediction performance on GSE19536 dataset [37]. Since miRNA-mRNA interactions are highly tissue-specific and development-specific, we focus on datasets from the same cancer type here. Although both datasets were from breast cancer samples, they used different microarray platforms for mRNA and miRNA profiling.

In order to perform a cross-database application of the model, we first find the mRNAs and miRNAs that are in common between the two studies. Since the studies use different microarray platforms with different probe IDs, we convert the mRNA probe IDs to their GeneBank accession numbers and the miRNA probe IDs to their miRBase IDs. This results in 14873 mRNAs and 232 miRNAs that are in common between the two studies.

The comparison of the heat maps generated from real and predicted data illustrates that SMLR is able to predict the overall expression profiles that reflect the ER status of the samples (See Fig. 4, top row). We observe the same behavior when the training and test datasets were switched (Fig. 4, bottom row). Taking the differentially expressed mRNAs from the predicted GSE22220 data (p-value < 0.01) and performing gene set enrichment, again finds functional annotations that are in better agreement with those obtained from the real data, when compared to the agreement of the annotations resulting from the TargetScan or negative correlation methods (Fig. 3-Right).

Although our main focus in this study is quantitative prediction of mRNA expression levels, some of the underlying predictors discovered by our model may be from direct miRNA-mRNA target interactions. Specifically, some of the coefficients w_i in Eq. "1" (which make up the matrix $\boldsymbol{B}$ in Eq. 2) may represent direct miRNA-mRNA targeting interactions. We assess the extent in which SMLR can discover such targeting interactions by comparing these interactions with known miRNA targets in miRTarBase and predicted targets in TargetScan.

The SMLR model was trained on both GSE22220 and GSE19536 datasets combined and the miRNA-mRNA pairs in the model with negative coefficients, representing a potential targeting effect, were collected. Here, we consider only the 248 miRNAs for which there was at least one such targeting interaction. There were on the average 8 experimentally validated targets for each of these miRNAs, listed in miRTarBase. TargetScan had an average of 341 predicted targets per miRNA. Considering miRTarBase as the ground truth, the accuracy of miRNA-mRNA target pairs predicted by SMLR was 0.10% (41 correct out of 40,633 predictions), whereas TargetScan had

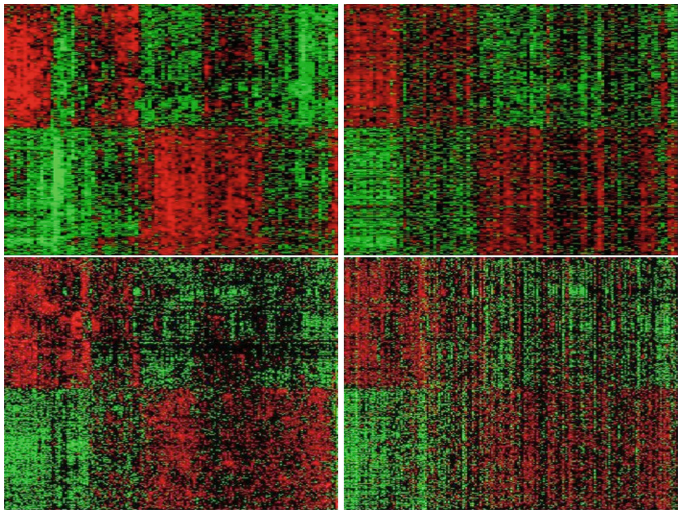

Fig. 4. Hierarchical clustering of true (left) and cross-database predicted (right) mRNA expression. Top: SMLR is trained with GSE22220 dataset and tested on GSE19536 (RMSE = 1.02). Bottom: SMLR is trained with GSE19536 dataset and tested on GSE22220 (RMSE = 1.26). Top 1000 most differentially expressed mRNAs with respect to ER-status are shown. Hierarchical clustering is only done on the real data (left); and the same row-column ordering is used to display the predicted data (right).

an accuracy of 1.12% (944 out of 84,489 predictions) and the negative correlation method had an accuracy of 0.05% (222 out of 428,048 predictions).

Although SMLR had a lower accuracy than TargetScan, we must note that the coverage of miRTarBase is currently very limited. Consequently, these accuracy measures are sensitive to availability of further experimentally validated target data. Furthermore, whereas SMLR finds interactions specific to the datasets it is trained with, namely the breast cancer samples, miRTarBase dataset and TargetScan predictions do not provide any context-specific information for their target interactions. Regardless of these drawbacks in the analysis, combining the predictions from SMLR and TargetScan, by intersecting their miRNA-mRNA target pair lists, achieves an accuracy of 2.17% (23 correct out of 1,060 common predictions), which is better than application of either method alone.

4 Discussion and Conclusion

In this study, we took a radically different approach to miRNA-mRNA interactions and used a multiple linear regression model to directly estimate the mRNA expression levels from miRNA data. Whereas traditional methods try to determine targets of individual miRNAs and rely on these target lists for downstream functional analysis, we estimate mRNA levels from the cellular context captured by the collection of miRNAs. The benefits and opportunities provided by our approach are tremendous. For

instance, our approach makes it possible to computationally predict mRNA levels for media, such as serum, where miRNAs are relatively stable and easy to extract and measure with current experimental techniques but mRNAs are less stable and more challenging to measure.

Traditionally, after identifying differentially regulated miRNAs, researchers would sift through hundreds or thousands of targets of these miRNAs and subjectively pick several targets of interest for further experimental validation, e.g., to test for binding of miRNA to mRNA or for differential regulation of the mRNA. Not only are these target lists non-specific to the tissue type, developmental stage, or environmental factors involved in an experimental study; they also ignore the fact that these genes are targeted by multiple miRNAs, some of which may not be differentially regulated or may be regulated in different directions. In our approach on the other hand, we build a model in a cell-type specific manner, connecting multiple miRNAs to each mRNA. We believe that a prioritization of the target genes based on estimated expression levels will result in a higher positive rate in validation experiments.

Our choice of the SMLR model for prediction of mRNA expression levels was based on its simplicity and interpretability. We believe that the linearity assumption used in SMLR provides an appropriate trade-off between the power and generality of the model and the number of parameters that can be correctly estimated from the currently available datasets. Furthermore, the interactions obtained from linear models were previously found to be better than those generated from Bayesian models and Neural Networks [33].

In this study, we mainly focused on breast cancer datasets and demonstrated that a model trained in one experimental platform can be successfully applied to miRNA data from an independent laboratory using different experimental platforms. Although it is possible to apply a model trained on one tissue type to miRNA data from another tissue type; the predicted gene expression values would not be as accurate as restricted the predictions to the same tissue and comparable experimental conditions. For example, applying the model trained on the breast cancer dataset GSE22220 to predict gene expression values from miRNA data in a prostate cancer study GSE20161 resulted in a mean squared error of 1.35, about 33% higher than the error when it was applied to another breast cancer dataset GSE19536. In our future work, we will build a repository of models for different tissue types and experimental conditions of interest. The limiting factor for building such a repository will be the availability of high quality paired miRNA and mRNA data collected from the same samples.

Although our main focus was not identification of the direct miRNA-mRNA targeting interactions, we show that the interactions with negative coefficients in our model can be indicative of direct regulation. Note that the targets from our model were generated only from the two breast cancer studies. We expect that a large scale modeling from all publicly available paired miRNA-mRNA datasets will provide target predictions that are in better agreement with experimentally validated targets. Motivated by the observation that targeting interactions obtained from two breast cancer datasets can improve the accuracy of TargetScan predictions, we expect that our approach will provide a means of improving sequence-based target predictions in a context-specific manner.

References

1. Bartel, D.P.: MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116**(2), 281–297 (2004)
2. Vasudevan, S., Tong, Y., Steitz, J.A.: Switching from repression to activation: microrna can up-regulate translation. Science **318**(5858), 1931–1934 (2007)
3. Hobert, O.: Gene regulation by transcription factors and microRNAs. Science **319**(5871), 1785–1786 (2008)
4. Fabian, M.R., Sonenberg, N., Filipowicz, W.: Regulation of mRNA translation and stability by microRNAs. Ann. Rev. Biochem. **79**(1), 351–379 (2010)
5. Huntzinger, E., Izaurralde, E.: Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat. Rev. Genet. **12**(2), 99–110 (2011)
6. Guo, H., et al.: Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature **466**(7308), 835–840 (2010)
7. Friedman, R.C., et al.: Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. **19**(1), 92–105 (2009)
8. Croce, C.M.: Causes and consequences of microRNA dysregulation in cancer. Nat. Rev. Genet. **10**(10), 704–714 (2009)
9. Lujambio, A., Lowe, S.W.: The microcosmos of cancer. Nature **482**(7385), 347–355 (2012)
10. Ørom, U.A., Lund, A.H.: Experimental identification of microRNA targets. Gene **451**(1–2), 1–5 (2010)
11. Vergoulis, T., et al.: TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res. **40**, D222–D229 (2011)
12. Hsu, S.D., et al.: miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. **42**(Database issue), D78–D85 (2014)
13. John, B., et al.: Human microRNA targets. PLoS Biol. **2**(11), e363 (2004)
14. Krek, A., et al.: Combinatorial microRNA target predictions. Nat. Genet. **37**(5), 495–500 (2005)
15. Lewis, B.P., Burge, C.B., Bartel, D.P.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell **120**(1), 15–20 (2005)
16. Maragkakis, M., et al.: DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res. **37**(Web Server issue), W273–W276 (2009)
17. Sethupathy, P., Megraw, M., Hatzigeorgiou, A.G.: A guide through present computational approaches for the identification of mammalian microRNA targets. Nat. Methods **3**(11), 881–886 (2006)
18. Pritchard, C.C., Cheng, H.H., Tewari, M.: MicroRNA profiling: approaches and considerations. Nat. Rev. Genet. **13**(5), 358–369 (2012)
19. Nam, S., et al.: miRGator: an integrated system for functional annotation of microRNAs. Nucleic Acids Res. **36**(Suppl. 1), D159–D164 (2008)
20. Huang, G.T., Athanassiou, C., Benos, P.V.: mirConnX: condition-specific mRNA-microRNA network integrator. Nucleic Acids Res. **39**(Suppl. 2), W416–W423 (2011)
21. Ritchie, W., Flamant, S., Rasko, J.E.J.: mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. Bioinformatics **26**(2), 223–227 (2010)
22. Peng, X., et al.: Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. BMC Genom. **10**(1), 373 (2009)

23. Sales, G., et al.: MAGIA, a web-based tool for miRNA and genes integrated analysis. Nucleic Acids Res. **38**(Suppl. 2), W352–W359 (2010)
24. Nam, S., et al.: MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. Nucleic Acids Res. **37**(Suppl. 2), W356–W362 (2009)
25. Kim, S., Choi, M., Cho, K.H.: Identifying the target mRNAs of microRNAs in colorectal cancer. Comput. Biol. Chem. **33**(1), 94–99 (2009)
26. Wang, H., Li, W.H.: Increasing MicroRNA target prediction confidence by the relative R(2) method. J. Theoret. Biol. **259**(4), 793–798 (2009)
27. Beck, D., et al.: Integrative analysis of next generation sequencing for small non-coding RNAs and transcriptional regulation in myelodysplastic syndromes. BMC Med. Genom. **4** (1), 19 (2011)
28. Huang, J.C., Morris, Q.D., Frey, B.J.: Bayesian inference of MicroRNA targets from sequence and expression data. J. Comput. Biol. **14**(5), 550–563 (2007)
29. Huang, J.C., Frey, B.J., Morris, Q.D.: Comparing sequence and expression for predicting microRNA targets using GenMiR3. In: Pacific Symposium on Biocomputing, pp. 52–63 (2008)
30. Su, N., et al.: Predicting microRNA targets by integrating sequence and expression data in cancer. In: 2011 IEEE International Conference on Systems Biology (ISB) (2011)
31. Stingo, F.C., et al.: A Bayesian graphical modeling approach to microRNA regulatory network inference. Ann. Appl. Stat. **4**(4), 2024–2048 (2010)
32. Liu, B., et al.: Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy. BMC Bioinform. **10**(1), 408 (2009)
33. Zhou, Y., Qureshi, R., Sacan, A.: Data simulation and regulatory network reconstruction from time-series microarray data using stepwise multiple linear regression. Netw. Model. Anal. Health Inform. Bioinform. **1**(1–2), 3–17 (2012)
34. Liu, B., Li, J., Cairns, M.J.: Identifying miRNAs, targets and functions. Briefings Bioinform. **15**(1), 1–19 (2014)
35. da Huang, W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. **4**(1), 44–57 (2009)
36. Enerly, E., et al.: miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. PLoS ONE **6**(2), e16915 (2011)
37. Buffa, F.M., et al.: microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res. **71**(17), 5635–5645 (2011)
38. Naume, B., et al.: Detection of isolated tumor cells in bone marrow in early-stage breast carcinoma patients: comparison with preoperative clinical parameters and primary tumor characteristics. Clin. Cancer Res. **7**(12), 4122–4129 (2001)