Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (B)

A Bayesian Framework to Improve MicroRNA Target Prediction by Incorporating External Information

Zixing Wang[1], Wenlong Xu[1], Haifeng Zhu[3] and Yin Liu[1,2]

[1]Department of Neurobiology and Anatomy, University of Texas Health Science Center at Houston, Houston, TX, USA. [2]University of Texas Graduate School of Biomedical Science, Houston, TX, USA. [3]Department of Melanoma Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA.

ABSTRACT: MicroRNAs (miRNAs) are small regulatory RNAs that play key gene-regulatory roles in diverse biological processes, particularly in cancer development. Therefore, inferring miRNA targets is an essential step to fully understanding the functional properties of miRNA actions in regulating tumorigenesis. Bayesian linear regression modeling has been proposed for identifying the interactions between miRNAs and mRNAs on the basis of the integrated sequence information and matched miRNA and mRNA expression data; however, this approach does not use the full spectrum of available features of putative miRNA targets. In this study, we integrated four important sequence and structural features of miRNA targeting with paired miRNA and mRNA expression data to improve miRNA-target prediction in a Bayesian framework. We have applied this approach to a gene-expression study of liver cancer patients and examined the posterior probability of each miRNA–mRNA interaction being functional in the development of liver cancer. Our method achieved better performance, in terms of the number of true targets identified, than did other methods.

KEYWORDS: microRNA target prediction, gene regulation, gene expression, prior information, sequence feature

SUPPLEMENT: Array Platform Modeling and Analysis (B)

CITATION: Wang et al. A Bayesian Framework to Improve MicroRNA Target Prediction by Incorporating External Information. *Cancer Informatics* 2014:13(S7) 19–25 doi: 10.4137/CIN.S16348.

RECEIVED: August 12, 2014. **RESUBMITTED:** October 14, 2014. **ACCEPTED FOR PUBLICATION:** October 16, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Methodology

FUNDING: This work is supported in part by NIH grant R01 LM010022 and the seed grant from the University of Texas Health Science Center at Houston. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Yin.Liu@uth.tmc.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

MicroRNAs (miRNAs) are highly conserved small RNAs that have diverse functions, including regulation of cellular differentiation, proliferation, and apoptosis.[1,2] These RNA molecules exert their function by inhibiting translation or inducing degradation of their target messenger RNAs. A given miRNA is able to pair with hundreds of transcripts by its seed miRNA nucleotides, allowing it to regulate complex gene-expression programs and induce global physiological changes.[3] Dysfunction of these miRNA molecules has been linked to several human diseases, including different types of cancer. Virtually, all examined tumor types have globally abnormal miRNA expression patterns, where miRNAs play regulatory roles as potential oncogenes or oncosuppressor genes.[4–6] Genome-wide profiling showed that about half of miRNA genes are localized in cancer-related genomic regions or fragile loci,[7] where mutations, deletions, or amplifications occur in many human tumors. These observations indicate that miRNAs are candidate genes for tumorigenesis and cancer progression. An essential step and major challenge to understanding the functions of miRNAs in cancer is identification of their target genes. Many computational and experimental approaches have been used to improve the reliability of miRNA-target prediction. TargetScan,[8–10] PicTar,[11] DIANA-microT,[12] miRanda,[13] and TargetS[14,15] are examples of computational approaches that are based on an analysis of miRNA and mRNA sequences. Generally, they use the following principles to predict miRNA targets:

1. Seed matches: the Waston-Crick pairing between the 5′ region of the miRNA centered on nucleotides 2–7 and the 3′ untranslated region (UTR) of the target mRNA.
2. Degree of conservation: a functional miRNA target is preferentially conserved across multiple species.
3. Thermodynamic stability, measured by the hybridization energy between miRNA and its candidate target site. It is believed that the total free energy of a functional targeting must be thermodynamically favorable, ie, negative valued.
4. Accessibility energy, which is the free energy required to unpair the nucleotides on the target site to make the target accessible to the miRNA.
5. Target site context, including local AU content; the target position within 3′ UTR; and the residue pairing at 3′ of the putative target site.[9]

These computational methods, which integrate multiple types of sequence and structural features, however, have low specificity and a high number of false positives for miRNA-target prediction. More importantly, predictions based on sequence and structural features only represent static miRNA–mRNA interactions. It is not clear to what extent these predicted interactions align with functional miRNA regulation in a particular phenotype or pathological condition. Thus, expression profiling has been proposed as an important information resource for discovering miRNA targets under different conditions. On the basis of this idea, some novel approaches have been developed to predict miRNA targets by integrating expression data into sequence-based prediction. Among them are GeneMiR++,[16,17] TaLasso,[18] HOCTAR,[19] BLasso,[20] MAGIA,[21] and HCTarget.[22] They mainly use paired miRNA and mRNA expression data from the same set of samples to refine the sequence-based prediction results and obtain more reliable miRNA targets. However, these approaches do not consider the full spectrum of available sequence and structural features of putative miRNA targets. Instead, they view all potential targets in sequence-based prediction results as equally biologically meaningful. Recently, Xu et al systematically evaluated the effects of sequence and structural features on miRNA-target prediction using the pSILAC dataset as a benchmark.[14,15] It was found that all these features were important for improving the accuracy of miRNA-target identification.

In this study, we combined the paired expression data of miRNAs and mRNAs from liver cancer patients, and the sequence and structural features of miRNA targeting to improve miRNA-target prediction. Our approach was based on a Bayesian linear regression model coupled with the Markov chain Monte Carlo (MCMC) algorithm. It uses both sequence and structural feature information to establish a prior probability of a miRNA-target interaction being functional, and paired miRNA and mRNA expression data to compute the likelihood of a putative miRNA-target interaction. By combining these two sources of information, our Bayesian method allows us to effectively sample from the large search space of putative miRNA–mRNA interactions and compute the posterior probability of each putative miRNA target. It represents a powerful means of reconstructing miRNA–mRNA interaction networks, specifically in liver cancer samples, and might help us uncover the mechanisms of tumorigenesis and progression in liver cancer.

Methods

Given a set of expression data of miRNA and mRNA, we modeled the interaction between miRNAs and target mRNAs using a linear model. The log-conditional likelihood function of data can be written in the following form, assuming a normal distribution:

$$p(y_g \mid X, \beta_g, \sigma_g^2) = N(X\beta_g, \sigma_g^2 I_N) \quad (1)$$

where $y_g \in R^N$ represents the collection of mRNA expression data with N number of samples. $X \in R^{N\times M}$ is the collection of miRNA expression data, where M is the number of miRNAs, and σ_g^2 is the noise. $\beta_g \in R^M$ is the regression coefficient vector of the gth mRNA.

Without additional sequence and structural feature information. The goal of this analysis is to identify a small subset of miRNA–mRNA interactions that are biologically meaningful. In the framework of variable selection,[23,24] an indicator matrix is defined as

$$r_{mg} = \begin{cases} 1 \; if \; \beta_{mg} \neq 0 \left(the \; m-th \; miRNA \; is \; selected \; for \; g-th \; mRNA \right) \\ 0 \; if \; \beta_{mg} = 0 \left(the \; m-th \; miRNA \; is \; not \; selected \; for \; g-th \; mRNA \right) \end{cases} \quad (2)$$

Here, r_{mg} is a binary indicator of whether the interaction between the mth miRNA and the gth mRNA is functional. In this model (without sequence and structural feature information), we only incorporated the computationally predicted, sequence-based miRNA-target information as prior information. We used an additional indicator matrix, C, in the current model, where the entry c_{mg} is an indicator whose value is 1 if the gth mRNA is a potential target of the mth miRNA in the database and 0 otherwise. We focused on the entries with $c_{mg} = 1$. We also assumed that r_{mg} is independent of each other and follows a Bernoulli distribution, as in the following equation:

$$p(r_{mg} \mid \pi) \sim \pi^{c} mg^{r} mg(1-\pi)^{c_{mg}(1-r_{mg})}, 0 \leq \pi \leq 1 \quad (3)$$

Here π can be regarded as the proportion of the true targets in databases.

We used a non-informative prior for β_g, σ_g^2:

$$p(\beta_g, \sigma_g^2) \sim \sigma_g^{-2} \quad (4)$$

The joint posterior distribution is written as

$$p(\beta_g,\sigma_g^2,r_{mg}\mid y_g,X,C,\pi) \sim \sigma_g^{-\left[\frac{N}{2}+1\right]}\times\exp\left[-\frac{1}{2\sigma_g^2}\left(y_g-Xc_{\cdot g}r_{\cdot g}\beta_g\right)^T\left(y_g-Xc_{\cdot g}r_{\cdot g}\beta_g\right)\right]\times\prod_{m=1}^{M}\pi^{c_{mg}r_{mg}}(1-\pi)^{c_{mg}(1-r_{mg})} \quad (5)$$

To efficiently search the parameter space of r_{mg} using MCMC sampling, we integrate β_g and σ_g^2 out; the marginal distribution of r_{mg} is proportional to

$$\frac{\Gamma\left(\frac{N-P}{2}\right)(N-p)^{\frac{p}{2}}\pi^{\frac{p}{2}}\left|s_g^2\left(X_{c_{\cdot g}r_{\cdot g}}^TX_{c_{\cdot g}r_{\cdot g}}\right)^{-1}\right|^{\frac{1}{2}}}{\Gamma\left(\frac{N}{2}\right)}\prod_{m=1}^{M}\pi^{c_{mg}r_{mg}}(1-\pi)^{c_{mg}(1-r_{mg})} \quad (6)$$

where p is the total number of miRNAs in the model, which is equal to $sum(c_{\cdot g}r_{\cdot g})$. Because of independence of r_{mg}, we can infer an individual r_{mg} conditional on r_{-mg}, where r_{-mg} is the vector of $r_{\cdot g}$ without the mth element and

$$p(r_{mg}\mid\pi,r_{-mg})\sim\frac{\Gamma\left(\frac{N-p}{2}\right)(N-p)^{\frac{p}{2}}\pi^{\frac{p}{2}}\mid S_g^2(X_{c_{\cdot g}r_{\cdot g}}^TX_{c_{\cdot g}r_{\cdot g}})^{-1}\mid^{1/2}}{\Gamma\left(\frac{N}{2}\right)} \pi^{c_{mg}r_{mg}}(1-\pi)^{c_{mg}(1-r_{mg})} \quad (7)$$

Because r_{mg} is binary, we can define its marginal distribution as a Bernoulli distribution with a success probability of

$$\lambda=\frac{p(r_{mg}=1\mid\pi,r_{-mg})}{p(r_{mg}=1\mid\pi,r_{-mg})+p(r_{mg}=0\mid\pi,r_{-mg})} \quad (8)$$

Here, we implemented a Gibbs sampler to sample each r_{mg}. We initialized the vector $r_{\cdot g}$ at random and then sampled each entry of r_{mg} with other entries r_{-mg} fixed on the basis of the Bernoulli distribution, with a success probability λ.

With additional sequence and structural features as prior information. To incorporate the sequence and structural features of miRNA-target sites into the model, we introduced an F-dimensional vector $f_{mg}=\left(f_{mg}^1,f_{mg}^2\ldots f_{mg}^F\right)$ that was composed of F features associated with each miRNA–mRNA pair (m,g). We denoted $\pi_{mg}=p(r_{mg}=1\mid c_{mg}=1,f_{mg},w)$ as the prior probability of $r_{mg}=1$ given F features. To simplify the model, we assumed that each of the F features independently contributes to π_{mg} with a weight of w_f, where $f=1,\ldots,F$. Here, w is an unknown parameter with positive values. We defined the prior as

$$\pi_{mg}=p(r_{mg}=1\mid c_{mg}=1,f_{mg},w)=\frac{1}{1+\exp(-w^Tf_{mg})} \quad (9)$$

We further specified a hyperprior on w as gamma distribution $w\sim Gamma(a,b)$, ensuring the positivity of the parameter. In this work, we included four types of features that play crucial roles in miRNA target recognition (Please see the Results section for the details of the features included in our model.). Therefore, the feature vector f has four dimensions, the same as w. These features should be normalized to obtain positive values that lie in the same range, with a bigger value corresponding to a higher prior probability.

Following the Gibbs sampling of r_{mg} given a success probability π_{mg}, we sampled w using Metropolis steps[25] so that we can update π_{mg}, depending on the value of the sequence features. The proposal is made via a truncated normal random walk kernel. The proposed w^{new} is then accepted with the probability

$$\min\left[\frac{\pi\left(r_{mg}\mid w_j^{new}\right)p\left(w_j^{new}\right)q\left(w_j^{old};w_j^{new}\right)}{\left(r_{mg}\mid w_j^{old}\right)p\left(w_j^{old}\right)q\left(w_j^{new};w_j^{old}\right)},1\right] \quad (10)$$

where $q\left(w_j^{old};w_j^{new}\right)$ is a truncated normal with mean w_j^{new} and is truncated at 0, given the positive nature of w. The variance of this distribution has to be set to accommodate an appropriate acceptance rate during MCMC sampling. The sampling of r_{mg} and w was iterated until the MCMC chain was converged. Using the MCMC sampling procedure, we could explore the search space and find the most relevant predictions using a stochastic search variable selection method. The posterior probability of an miRNA–mRNA interaction, that is, $p(r_{mg}=1\mid c_{mg}=1,Y,X)$, can be estimated directly from the MCMC sampling results by taking the proportion of MCMC iterations for which $r_{mg}=1$.

Results

We studied the regulatory roles of miRNAs in a dataset of matched miRNA and mRNA expression data for 125 patients with liver cancer from The Cancer Genome Atlas (TCGA). We log-transformed the expression data to ensure that they approximately followed normal distribution during the data preprocessing step. The computationally predicted miRNA–mRNA interactions were extracted from TargetScanHuman (release 6.1)[10] and mapped to the expression dataset.[10] This yielded 67,798 interactions between 170 human miRNAs and 4973 mRNA transcripts, which were used as the prior information for our first model without the addition of miRNA features.

To determine the effects of the additional sequence and structural features of miRNA on target prediction, we obtained context scores and aggregated probability of conserved targeting scores for each miRNA–mRNA pair from TargetScanHuman. The probability of conserved targeting score is a target site conservation score and has been used to measure the degree of miRNA target sequence conservation across multiple species. We also calculated the thermodynamic stability (ΔG) and the accessibility energy ($\Delta\Delta G$) for each putative miRNA–mRNA

interaction.[14,15] Therefore, totally four sequence and structural features were integrated into our model to establish the prior probability of a miRNA-target interaction being biologically functional. We then compared this algorithm to the method without the addition of these four features. The miRNA-target interaction set with the highest scores from each approach was selected and compared in terms of its enrichment results in an experimentally validated interaction.

In our Bayesian framework without additional miRNA features, the parameter π of the Bernoulli distribution reflects the prior belief about the proportion of true targets in the computationally predicted miRNA–mRNA interactions. Since there were 67,798 putative miRNA–mRNA interactions included in the liver cancer dataset, we set $\pi = 0.07$, indicating that 7% of putative interactions are true; thus, the expected number of miRNA–mRNA interactions for each mRNA would be approximately equal to 1. In our model with the miRNA sequence and structural features, we tuned the prior probability of each interaction according to the values of their corresponding features. We set the hyperparameters for the gamma distribution of weights as $a = 1.5$ and $b = 0.05$, and the variance of the truncated normal proposal distribution of w_j to 0.01 so that we could obtain an acceptance rate close to 25%. Figure 1 shows the summary trace plots for the number of r_{gm} samplings and the corresponding log-posterior probabilities for our two models (with and without the additional features). In this application, the MCMC chain was run for 10^6 iterations, starting from a randomly chosen set of 5000 miRNA–mRNA interactions, so that each gene was targeted by approximately one miRNA on average, which is consistent with our prior specification.

To assess the performance of our approach, we evaluated the enrichment scores of the results from experimentally validated miRNA–mRNA interactions. If the top-ranked miRNA and mRNA interactions identified from an algorithm include more experimentally validated targets, this algorithm will be considered to have better performance because more predicted interactions can be validated. Here, we extracted the experimentally validated target information from TarBase 6.0, which includes more than 65,000 manually

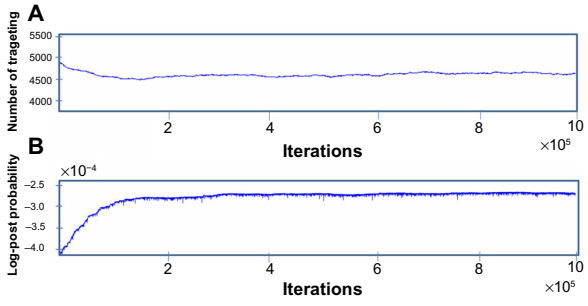

Figure 1. Trace plots for (**A**) the number of selected miRNA–mRNA interactions and (**B**) the log-posterior probability along the number of iterations.

curated, experimentally validated miRNA-gene interactions from eight species.[26] To examine the overlaps between the TarBase information and our prediction results, we mapped all miRNAs in our dataset to the miRNA families in TarBase using the annotations in miRBase. In the liver cancer expression dataset, 609 miRNA–mRNA interactions have been biologically verified. We found that 68 and 79 of these interactions, without and with the addition of miRNA features, respectively, overlapped with the top 5000 targets detected by our model; the well-known GenMiR++ method only identified 56 interactions. On the basis of this observation, we obtained the numbers of false positives and false negatives, and calculated the corresponding statistical significance of the number of true targets identified by different methods using the hypergeometric distribution. For a given number of identified true targets, the smaller the *P*-value, the more enriched the predicted set of targets in the experimentally validated interaction. The results demonstrate that our model with the addition of miRNA sequence and structural features resulted in a most significant *P*-value, compared to the non-feature model and the GenMiR++ method, as shown in Table 1. We also examined the top 500 targets and observed similar results. The experimentally validated targets that were predicted by at least two of the three methods are listed in Table 2.

To further investigate the function of our predicted targets and the potential regulatory roles of miRNA in patients with liver cancer, we analyzed the biological relevance of the target genes in a KEGG pathway enrichment study.[27] We used the KEGG pathway annotation to measure the enrichment of the top 200 genes predicted by different methods using the GeneCodis 3.0 tool. As a result, several KEGG pathways were found to be significantly enriched in the results obtained from our models and GenMiR++ (Table 3). Both of our models (with and without additional features) resulted in significantly enriched pathways related to cancer and focal adhesion. For GeneMiR++, the two most prominent pathways were related to cell cycle and focal adhesion.

Among the miRNAs shown in Table 2, hsa-miR-145 and hsa-miR-21 are key regulators during hepatocellular carcinoma genesis.[28,29] In particular, hsa-miR-145 functions as a tumor suppressor in liver cancer by targeting the chromatin modification enzyme, histone deacetylase.[28] In this study, we discovered many novel functional targets of hsa-miR-145, including MUC1. We used the pair hsa-miR-145-MUC1 to illustrate the effectiveness of our model. We grouped liver cancer patients by their hsa-miR-145 expression level (higher or lower than average). The patients in the high hsa-miR-145 group had significantly lower MUC1 expression than those in the other group (the *P*-value was 0.06, one-sided Wilcoxon test). Their cumulative distributions displayed a negative shift of MUC1 (Fig. 2). This example further confirms the gene down-regulatory effect of hsa-miR-145 and indicates that MUC1 is a reliable target gene. It is expected that the hsa-miR-145-MUC1 pair will provide

Table 1. Enrichment values of experimentally validated targets obtained from our feature-dependent model (feature-MCMC), the model without additional features (non-feature), and the GenMiR++ method. In the top 5000 or 500 predicted interactions, the numbers of experimentally validated targets (true positives), false positives, false negatives, and enrichment significance (*P*-value) are given. *P*-values were calculated on the basis of the hypergeometric distribution.

MODEL	TOP	TRUE POSITIVE	FALSE POSITIVE	FALSE NEGATIVES	*P*-VALUE
Feature-MCMC	5000	79	4921	530	2.23E-06
	500	15	485	594	7.90E-05
Non-feature	5000	68	4932	521	1.51E-04
	500	15	485	594	7.90E-05
GenMiR++	5000	56	4944	553	2.31E-02
	500	10	490	599	5.41E-03

Table 2. Predicted experimentally validated targets obtained from our feature-dependent model (feature-MCMC), the model without additional features (non-feature), and the GenMiR++ method.

miRNA	GENE	FEATURE-MCMC	NON-FEATURE	GenMiR++
hsa-miR-103a-3p	Smarce1	×		×
hsa-miR-103a-3p	FKBP1A	×		×
hsa-miR-103a-3p	BCKDK	×	×	
hsa-miR-103a-3p	CCNE1	×		×
hsa-miR-103a-3p	aadat	×	×	
hsa-miR-103a-3p	SCAF1	×		×
hsa-miR-10a-5p	NDUFB6	×	×	
hsa-miR-145–5p	aph1a	×		×
hsa-miR-145–5p	MUC1	×	×	
hsa-miR-16–5p	CCNE1	×	×	×
hsa-miR-16–5p	Tppp3	×	×	
hsa-miR-185–5p	CCNE1	×		×
hsa-miR-186–5p	TMEM183A	×	×	
hsa-miR-191–5p	Mpst	×	×	
hsa-miR-19b-3p	WBP2	×	×	
hsa-miR-21–5p	TPM1	×	×	×
hsa-miR-22–3p	BTF3L1	×		×
hsa-miR-24–3p	vps25	×	×	
hsa-miR-24–3p	MARCKSL1	×	×	
hsa-miR-29a-3p	DNMT3A	×		×
hsa-miR-32–5p	Hivep1	×	×	
hsa-miR-32–5p	BCAT2	×		×
hsa-miR-34a-5p	MAGEA12	×	×	×
hsa-miR-34a-5p	Magea6	×	×	×
hsa-miR-7–5p	TCOF1	×		×
hsa-miR-7–5p	Pole4	×	×	
hsa-miR-7–5p	c18orf10	×	×	
hsa-miR-7–5p	dtymk	×	×	×
hsa-miR-93–5p	Gramd1a	×		×

Table 3 Top 10 enriched KEGG pathways from our feature-dependent model (feature-MCMC), the model without additional features (non-feature), and the GenMiR++ method.

FEATURE-MCMC MODEL	NUMBER OF GENES	*P*-VALUE
Pathways in cancer	14	5.32E-07
Focal adhesion	10	8.61E-06
Regulation of actin cytoskeleton	9	5.32E-05
Leukocyte transendothelial migration	6	0.0003
Pathways in cancer, focal adhesion	6	0.0001
Leukocyte transendothelial migration, adhere junction	4	4.04E-05
Adhere junction, bacterial invasion of epithelial cells	4	3.94E-05
Leukocyte transendothelial migration, adhere junction, tight junction	3	0.0002
Long-term depression, progesterone-mediated oocyte maturation	3	0.0002
Regulation of actin cytoskeleton, focal adhesion, leukocyte transendothelial migration, adhere junction	3	0.0001
NON-FEATURE MODEL	**NUMBER OF GENES**	***P*-VALUE**
Pathways in cancer	8	0.0001
Focal adhesion	8	2.27E-05
Regulation of actin cytoskeleton	7	0.00022
Huntington's disease	6	0.0002
Pathways in cancer, focal adhesion	5	0.0001
Regulation of actin cytoskeleton, focal adhesion	5	0.0001
Pathway in cancer, focal adhesion, small cell lung cancer	4	0.0001
Pathway in cancer, focal adhesion, ECM-receptor interaction	3	0.0003
Regulation of actin cytoskeleton, focal adhesion, leukocyte transendothelial migration, bacterial invasion of epithelial cells	3	0.0001
Adhere junction, bacterial invasion of epithelial cells		0.0001
GenMiR++	**NUMBER OF GENES**	***P*-VALUE**
Cell cycle	10	6.41E-08
Focal adhesion	7	0.0003
Pyrimidine metabolism	6	8.34E-05
Focal adhesion, amoebiasis	6	1.83E-06
Pathways in cancer, small cell lung cancer	5	0.0003
Focal adhesion, ECM-receptor interaction	5	3.58E-05
DNA replication	4	0.0002
Pathways in cancer, focal adhesion, amoebiasis	4	9.02E-05
Focal adhesion, ECM-receptor interaction, amoebiasis	4	9.02E-05
DNA replication, cell cycle	3	5.44E-05

novel hypotheses for testing the roles of MUC1 in liver cancer development. If successful, it can serve as a biomarker for better directing the diagnosis and treatment of liver cancer patients.

Conclusions

In this study, we integrated matched miRNA and mRNA expression data with the sequence and structural features of miRNA seeds to improve miRNA target prediction. Compared to previous approaches,[25] our model restricts our search space to the putative miRNA targets obtained from a well-known miRNA target prediction database; thus, our model led to significantly less computational complexity but higher target prediction specificity. In addition, using a Bayesian linear regression model, we successfully incorporated four key features of miRNA–mRNA interactions; each assigned a different weight by the MCMC sampling procedure, as the prior knowledge in our model. Our investigation of paired miRNA and mRNA expression profiles in liver cancer patients successfully demonstrated the advantages of our feature-dependent model. Our results showed that the top interactions identified by our feature-dependent model are significantly more enriched in experimentally validated targets and are more biologically meaningful than are those identified by the GenMiR++ method or the model without additional feature information.

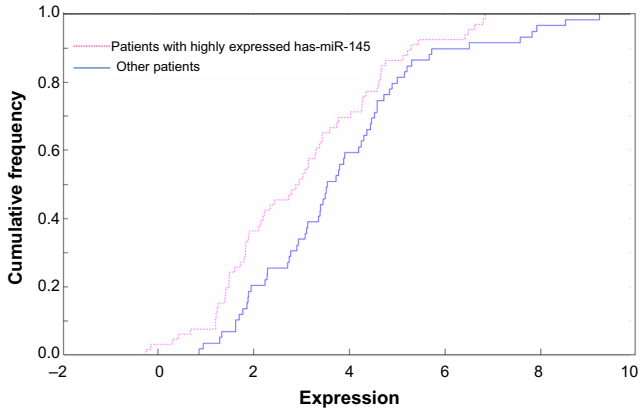

Figure 2. Down-regulatory effect of hsa-miR-154 on MUC1.The liver cancer patients were grouped according to their hsa-miR-145 expression levels (higher or lower than average).The cumulative distribution of the MUC1 expression levels in these two groups of patients was plotted, respectively (high hsa-miR-154, red dashed line; low hsa-miR-154, blue solid line). The *x*-axis represents the MUC1 expression levels represented by the Reads Per Kilobase of transcript per Million mapped reads (RPKM) values from the RNA-seq data.

In addition, with the recent intensive research in this field, a large body of experimentally verified miRNA target information has accumulated in available databases, such as StarBase[30] and miRWalk.[31] There is a strong interest in leveraging this information to improve target prediction sensitivity and accuracy, and this will be the focus of our future work. From a Bayesian perspective, we expect to be able to easily incorporate this information by assigning different prior distributions to the information sources according to their reliability, as in a proposed prior Lasso framework.[32]

Author Contributions

Conceived and designed the experiments: ZW, YL. Analyzed the data: ZW, WX, HZ. Wrote the first draft of the manuscript: ZW, YL. Contributed to the writing of the manuscript: ZW, YL. Agree with manuscript results and conclusions: ZW, XW, HZ, YL. Jointly developed the structure and arguments for the paper: ZW, YL. Made critical revisions and approved final version: ZW, XW, HZ, YL. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 2004;116(2):281–97.
2. Lee G, Sehgal R, Wang Z, et al. Essential role of grim-led programmed cell death for the establishment of corazonin-producing peptidergic nervous system during embryogenesis and metamorphosis in Drosophila melanogaster. *Biol Open*. 2013;2(3):283–94.
3. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature*. 2008;455(7209):58–63.
4. Chen D, Sun Y, Yuan Y, et al. miR-100 induces epithelial-mesenchymal transition but suppresses tumorigenesis, migration and invasion. *PLoS Genet*. 2014;10(2):e1004177.
5. Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol Med*. 2012;4(3):143–59.
6. Lee G, Wang Z, Sehgal R, et al. Drosophila caspases involved in developmentally regulated programmed cell death of peptidergic neurons during early metamorphosis. *J Comp Neurol*. 2011;519(1):34–48.
7. Calin GA, Sevignani C, Dumitru CD, et al. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci USA*. 2004;101(9):2999–3004.
8. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*. 2009;19(1):92–105.
9. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*. 2007;27(1):91–105.
10. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 2005;120(1):15–20.
11. Krek A, Grün D, Poy MN, et al. Combinatorial microRNA target predictions. *Nat Genet*. 2005;37(5):495–500.
12. Maragkakis M, Alexiou P, Papadopoulos GL, et al. Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*. 2009;10:295.
13. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. *Nucleic Acids Res*. 2008;36(Database issue):D149–53.
14. Xu W, San Lucas A, Wang Z, Liu Y. Identifying microRNA targets in different gene regions. *BMC Bioinformatics*. 2014;15(S4):11.
15. Xu W, Wang Z, Liu Y. The characterization of microRNA-mediated gene regulation as impacted by both target site location and seed match type. *PLoS One*. 2014;9(9):e108260.
16. Huang JC, Babak T, Corson TW, et al. Using expression profiling data to identify human microRNA targets. *Nat Methods*. 2007;4(12):1045–9.
17. Huang JC, Morris QD, Frey BJ. Bayesian inference of MicroRNA targets from sequence and expression data. *J Comput Biol*. 2007;14(5):550–63.
18. Muniategui A, Nogales-Cadenas R, Vázquez M, et al. Quantification of miRNA-mRNA interactions. *PLoS One*. 2012;7(2):e30766.
19. Gennarino VA, Sardiello M, Avellino R, et al. MicroRNA target prediction by expression analysis of host genes. *Genome Res*. 2009;19(3):481–90.
20. Zhong M, Liu R, Liu B. Bayesian analysis for miRNA and mRNA interactions using expression data. arXiv preprint arXiv. 2012;1210.3456:24.
21. Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C. MAGIA, a web-based tool for miRNA and genes integrated analysis. *Nucleic Acids Res*. 2010;38(Web Server issue):W352–59.
22. Su N, Qian M, Deng M. Integrative approaches for microRNA target prediction: combining sequence information and the paired mRNA and miRNA expression profiles. *Curr Bioinform*. 2013;8(1):9.
23. Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. *Bioinformatics*. 2003;19(1):90–7.
24. Wang Z, San Lucas FA, Qiu P, Liu Y. Improving the sensitivity of sample clustering by leveraging gene co-expression networks in variable selection. *BMC Bioinformatics*. 2014;15:153.
25. Stingo FC, Chen YA, Vannucci M, Barrier M, Mirkes PE. A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann Appl Stat*. 2010;4(4):2024–48.
26. Vergoulis T, Vlachos IS, Alexiou P, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*. 2012;40(Database issue):D222–29.
27. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res*. 2012;40(Web Server issue):W478–83.
28. Noh JH, Chang YG, Kim MG, et al. MiR-145 functions as a tumor suppressor by directly targeting histone deacetylase 2 in liver cancer. *Cancer Lett*. 2013;335(2):455–62.
29. Yuan SF, Li KZ, Wang L, et al. Expression of MUC1 and its significance in hepatocellular and cholangiocarcinoma tissue. *World J Gastroenterol*. 2005;11(30):4661–6.
30. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*. 2014;42(Database issue):D92–7.
31. Dweep H, Sticht C, Pandey P, Gretz N. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform*. 2011;44(5):839–47.
32. Wang Z, Xu W, Anthony San Lucas F, Liu Y. Incorporating prior knowledge into gene network study. *Bioinformatics*. 2013;29(20):2633–40.