Bioinformatics, 31(21), 2015, 3529–3536
doi: 10.1093/bioinformatics/btv392
Advance Access Publication Date: 30 June 2015
Original Paper

Systems biology

Integrating full spectrum of sequence features into predicting functional microRNA–mRNA interactions

Zixing Wang[1], Wenlong Xu[1] and Yin Liu[1,2,*]

[1]Department of Neurobiology and Anatomy, University of Texas Health Science Center at Houston and [2]University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 23, 2014; revised on May 31, 2015; accepted on June 23, 2015

Abstract

Motivation: MicroRNAs (miRNAs) play important roles in general biological processes and diseases pathogenesis. Identifying miRNA target genes is an essential step to fully understand the regulatory effects of miRNAs. Many computational methods based on the sequence complementary rules and the miRNA and mRNA expression profiles have been developed for this purpose. It is noted that there have been many sequence features of miRNA targets available, including the context features of the target sites, the thermodynamic stability and the accessibility energy for miRNA-mRNA interaction. However, most of current computational methods that combine sequence and expression information do not effectively integrate full spectrum of these features; instead, they perceive putative miRNA–mRNA interactions from sequence-based prediction as equally meaningful. Therefore, these sequence features have not been fully utilized for improving miRNA target prediction.

Results: We propose a novel regularized regression approach that is based on the adaptive Lasso procedure for detecting functional miRNA–mRNA interactions. Our method fully takes into account the gene sequence features and the miRNA and mRNA expression profiles. Given a set of sequence features for each putative miRNA–mRNA interaction and their expression values, our model quantifies the down-regulation effect of each miRNA on its targets while simultaneously estimating the contribution of each sequence feature to predicting functional miRNA–mRNA interactions. By applying our model to the expression datasets from two cancer studies, we have demonstrated our prediction results have achieved better sensitivity and specificity and are more biologically meaningful compared with those based on other methods.

Availability and implementation: The source code is available at: http://nba.uth.tmc.edu/homepage/liu/miRNALasso.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: Yin.Liu@uth.tmc.edu

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNA molecules (containing about 22 nucleotides) that play important gene-regulatory roles in plants and animals (Bartel, 2004). These small RNA molecules exert their function through interacting with their mRNA targets. The base paring of the miRNA with the 3′-untranslated region (3′-UTR) of their targets triggers the mRNA degradation and/or protein translational inhibition (Fabian *et al.*, 2010; Huntzinger

and Izaurralde, 2011). Since their initial discovery, miRNAs have emerged as key gene regulators in a variety of fundamentally important biological processes, including cell proliferation and apoptosis (Iorio and Croce, 2012; Lee *et al.*, 2011). Dysregulation of miRNA function has played a role in the pathophysiology of many human diseases, particularly in cancer (Calin *et al.*, 2004; Chen *et al.*, 2014).

MiRNAs represent one of the most abundant classes of gene-regulatory molecules in mammals. There are more than 2000 distinct miRNAs identified in human (Kozomara and Griffiths-Jones, 2011). It is also demonstrated that >60% of human mRNAs are subject to miRNA-mediated post-transcriptional regulation (Friedman *et al.*, 2009). Among the miRNA–mRNA regulatory relationships, multiple miRNAs are often required to act cooperatively to target the same gene. On the other hand, a given miRNA is able to pair with up to hundreds of genes (Selbach *et al.*, 2008). This multiplicity effect poses a great challenge to comprehensively depict a functional miRNA–mRNA regulatory network. Over the past decade, many miRNA algorithms and tools have been independently developed to study the miRNA–mRNA interactions, including TargetScan (Friedman *et al.*, 2009; Garcia *et al.*, 2011; Grimson *et al.*, 2007; Lewis *et al.*, 2005), PicTar (Krek *et al.*, 2005), MicroT_CDS (Maragkakis *et al.*, 2009), miRanda (Betel *et al.*, 2008), RNAhybrid (Kruger and Rehmsmeier, 2006), and TargetS (Xu *et al.*, 2014a). The majority of these algorithms are based on several well-known rules derived from important sequence features of target recognition, including sequence complementarity (seed matching), thermodynamic stability, target site context, target site accessibility and the degree of site conservation (Friedman *et al.*, 2009; Grimson *et al.*, 2007; Kertesz *et al.*, 2007; Lewis *et al.*, 2005). Despite these efforts, the number of false predictions obtained from these computational methods is still high, and more importantly, these predictions only represent static miRNA–mRNA interactions. Therefore, expression profiling has been proposed as a complementary information resource to study dynamic miRNA–mRNA regulatory relationship. As a result, many approaches tend to integrate this expression information into sequence-based target predictions to obtain more reliable miRNA targets. These methods assume that the expression levels of a given miRNA and its targets are inversely correlated due to the down-regulation effect of miRNAs on its targets. For example, GenMiR++ computes the probabilities of having an interaction between a miRNA and its putative targets in a Bayesian framework (Huang *et al.*, 2007). A Lasso-based method enforces a sparseness solution and combines the RISC availability with sequenced-based prediction and expression data (Lu *et al.*, 2011). Another related method named TaLasso includes non-positive constraints in their regularized least squares (Muniategui *et al.*, 2012). However, when integrating the sequence-based predictions on miRNA–mRNA interactions with the expression data, these approaches perceive these putative interactions as equally meaningful, as long as they meet certain criteria defined by the sequence-based prediction rules. Therefore, these approaches neglect the full spectrum of available sequence features of miRNA targets. A recent study addresses this problem to some degree by combining sequence context scores and conservation information with expression profiles. This model is specifically designed for miRNA-overexpression experiments to identify targets of a particular miRNA in different cell conditions (Li *et al.*, 2014).

In this article, we propose a novel regularized regression approach for detecting miRNA–mRNA interactions. Our approach takes into account the miRNA and mRNA expression profiles from matched samples while incorporating sequence features of putative targets. Beside the seed match rule between a miRNA and its targets, we use three additional features: (i) Thermodynamic stability, measured by the hybridization energy between a miRNA and its candidate target sites. For a successful hybridization, the binding energy ΔG_{hybrid} must be thermodynamically favorable, i.e. negatively valued (Rehmsmeier *et al.*, 2004); (ii) Target site accessibility, which is energetically measured by $\Delta\Delta G$, the free energy required to unpair the nucleotides on the target site to make the target accessible. It includes two components: ΔG_{hybrid} and $\Delta G_{disruption}$. The latter is the cost of altering the local structure of the mRNAs, which also prefers to be negatively valued (Kertesz *et al.*, 2007); (iii) Target site context. Several context features are related to miRNA targeting and have been found to boost the target site efficacy (Garcia *et al.*, 2011; Grimson *et al.*, 2007). Among these features, the local AU content near the target site, the target position within 3' UTR, the residue pairing at 3' end of miRNA, the target-site abundance, and the seed-pairing stability are used as the determinants to generate the target site context score by the TargetScan algorithm (Garcia *et al.*, 2011; Grimson *et al.*, 2007). For different types of target sites, the contribution of these context features are calculated using different sets of regression parameters deduced by the algorithm. The combined context value, named the Context Score is then calculated as the sum of the contribution of these context features, plus the mean repression level of the targets with the corresponding seed match type. To ease the implementation of our method, we use the context score directly as an additional sequence feature. Our algorithm only quantifies the miRNA–mRNA interactions from an initial set of putative miRNA–mRNA pairs that were obtained from TargetScanHuman. Therefore, it works as a filter to the TargetScanHuman prediction results, and is expected to achieve a better accuracy by combining these sequence features and the expression data of miRNAs and mRNAs together. Through real data application, we have demonstrated that our method has achieved better sensitivity and specificity in predicting miRNA-target interactions, and the predicted targets are more biologically meaningful than other methods we compared.

2 Methods

2.1 Mathematical model

Given an expression dataset corresponding to totally G mRNAs and M miRNAs in N samples, we let vectors $x_j = [x_{j1}, x_{j2}, \ldots, x_{jN}]$ and $z_k = [z_{k1}, z_{k2}, \ldots, z_{kN}]$ represent the expression levels of mRNA j and miRNA k, in samples 1 to N, respectively. Let c_{jk} be an indicator variable whose value is 1 if mRNA j is a putative target of miRNA k predicted in existing sequence databases and 0 otherwise. Then, a linear relationship between the expressions of each mRNA j and the K miRNAs is assumed and it is represented by the following linear model,

$$x_j = \sum_{k=1}^{K} \beta_{jk} . c_{jk} . z_k + \varepsilon_j \quad (1)$$

where ε_j is a standard Gaussian noise, and β_{jk} represents the degree of regulatory effect for each miRNA–mRNA interaction. Considering the down-regulation effect of miRNA, β_{jk} is preferentially a negative value. We further assume that x_j are standardized and consider a model without intercept. Within this context, the goal of study is to find a subset of miRNAs that regulate each mRNA with $\beta < 0$.The number of miRNAs that putatively regulate an mRNA is often larger than the number of samples, which possibly results in a high-dimensional model selection problem. To ensure proper estimation of parameter β, a L1 penalized least square estimator, known as the Lasso technique has been widely adapted as

a model selection method (Cao *et al.*, 2014; Lu *et al.*, 2011; Wang *et al.*, 2013). With the nature of sparse solution of Lasso, most of these regression coefficients β shrink to 0 and the nonzero ones will be fully recovered. Therefore, an L1 norm constraint on the coefficients is enforced in the model as the following:

$$\min\left\{\frac{1}{2}\|x_j - \sum_{k=1}^{K}\beta_{jk}.c_{jk}.z_k\|_2 + \lambda_j\sum_{k=1}^{K}|\delta_{jk}.\ \beta_{\ jk}.c_{jk}|\right\} \quad (2)$$

subject to $\beta_{jk} \leq 0$, for $k = 1,2,\ldots K$, where λ_j determines the degree of regularization of nonzero β_{jk}, and K is the total number of miRNAs in the dataset. The scale parameters $\delta_{jk} \in [0, 1]$ are usually fixed as unit one. Non-positivity constraints are added to ensure that the solution includes only negative relationships between mRNA and miRNA expressions. This is a convex problem and thus, the local minimum found by the algorithm is also guaranteed to be the global minimum.

To incorporate the prior knowledge of sequence features, we propose to automatically learn the scale parameter δ from data. We define δ_{jk} as a weighted sum of sequence features on the interaction between the j^{th} mRNA and the k^{th} miRNA. i.e.

$$\delta_{jk} = \sum_{t}\omega_t f_t^{jk} \quad (3)$$

where f_t^{jk} is the t^{th} feature for the interaction between mRNA j and miRNA k. Because we are interested in the relative contributions of different features, we further add the constraints that $\sum_t \omega_t = 1$ with constraint on the weights of each feature $\omega_t \geq 0$.

In an adaptive Lasso solution, the scale parameter δ_{jk} further refines the degree of regularization of nonzero β_{jk} according to the prior knowledge of each putative miRNA–mRNA interaction. Recall that the Lasso estimates can be interpreted as maximum a posteriori (MAP) estimates under Laplace priors (Tibshirani, 1996; Wang *et al.*, 2013), similarly, we compute the MAP estimates of β under the adaptive scale parameter in a Bayesian framework. Specifically, the conditional probability of β given the scale parameter has the following form:

$$P(\beta_{jk}|\delta_{jk}, \lambda_j) = \frac{1}{Z(\lambda_j\delta_{jk})}\exp(-\lambda_j\delta_{jk}|\beta_{jk}|) \quad (4)$$

It refers to a Laplacian prior with mean equal to 0 and the scale parameter related to δ_{jk}. $Z(\lambda_j\delta_{jk})$ is a normalization factor and it can be easily derived as the exponential integral equal to $2/\lambda_j\delta_{jk}$(Lee *et al.*, 2010). Our objective is to compute the MAP estimate of β and to simultaneously estimate the feature weights ω with the following optimization problem

$$\min\left\{\frac{1}{2}\|x_j - \sum_{k=1}^{K}\beta_{jk}.c_{jk}.z_k\|_2 + \lambda_j\sum_{k=1}^{K}|\delta_{jk}.\beta_{jk}.c_{jk}| + \sum_{k=1}^{K}\log(2/\lambda_j\delta_{jk})\right\} \quad (5)$$

The model can be illustrated as a network in Figure 1. This optimization problem can be solved by an iterative algorithm which (i) minimizes the equation over β by fixing ω; and (ii) minimizes the equation over ω by fixing β iteratively until convergence (Lee *et al.*, 2010). Both sub-problems are convex and they can be solved efficiently via an interior-point method and a gradient descent method, respectively. When compared with other convex quadratic programs, the interior-point method we implement here uses a preconditioned conjugate gradients algorithm to compute the search direction, which lends itself to efficiently solve large-scale L1-regularized least-squares problems (Kim *et al.*, 2007).

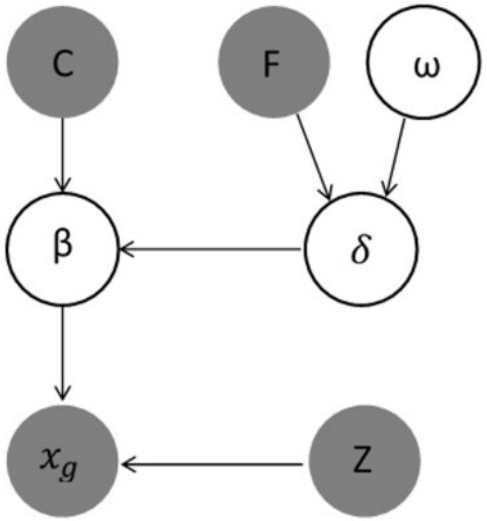

Fig. 1. Graphical model representation of adaptive Lasso. Shaded nodes denote known variables and unshaded ones are unknown parameters

2.2 Selection of the L1-penalty parameter

In the Lasso framework, the regularization parameter λ controls the sparsity of the selected model and its possible value often falls in the interval $[0, \lambda_j^{max}]$, where

$$\lambda_j^{max} = 2\cdot\max_k|(x_j\cdot z_k^T)| \quad (6)$$

However, due to the scale parameter δ in the adaptive Lasso, λ_j^{max} is usually greater than the predefined value in lasso. The optimal value of λ_j can be selected based on cross-validation. There are two ways to select λ_j: the global and the local parameter tuning methods. The global tuning method first identifies the maximum value of the L1 penalty where the optimal solution is a null vector for all the mRNAs, and then decreases the penalty gradually by searching through all the possible solution space, so it selects a general regularization parameter for all the mRNAs. For the local tuning method, it computes λ_j independently for each mRNA. It has been noted that the global tuning method performed better in selecting miRNA–mRNA interactions from the whole set of putative interactions than the local tuning method (Muniategui *et al.*, 2012), so the global tuning method is chosen in our implementation.

2.3 Pre-processing of sequence features

As previously stated, we extract three sequence features for each of the putative miRNA–mRNA pairs. We obtain the context score of target sites from TargetScanHuman (release 6.1). We also calculate the total hybridization energy (ΔG) and accessibility energy ($\Delta\Delta G$) for each miRNA target pairing. All of these three features yield negative values when functional miRNA targeting occurs. The smaller the value obtained from the formation of miRNA-target duplex, the more likely it suggests a functional target miRNA binding. To accommodate to our adaptive Lasso framework, we transform each of the features to a positive value in the following form:

$$f_t^{jk}(\text{tran}) = \text{sigmoid}\,(b_t * f_t^{jk}) = \frac{1}{1+\exp(-b_t * f_t^{jk})}, \quad (7)$$

where f_t^{jk} represents the t^{th} feature for the mRNA j and miRNA k pair. In this data pre-processing step, we include an extra parameter, b_t, where b_t is a scale factor for the t^{th} feature and is larger than

zero. Therefore, the transformed features will keep their original rankings and fall into the same range (0, 1/2). When the scale parameter approaches to infinity, the transformed feature values tend to be indistinguishable and will approach to 0. In this case, we would lose the informativeness of sequence features to the scale parameter. The same problem occurs if the parameter b is too small, where all transformed feature values will be close to 1/2. Therefore, an appropriate value of b would be critical for variable selection in the adaptive Lasso procedure. In the practical implementation of this study, the parameter is tuned for each feature set individually to achieve an optimal performance of the model.

2.4 Expression data collection

The first dataset, referred as Multi Class Cancer (MCC) dataset, has been used to evaluate both the GenMiR++ and the TaLasso methods. The dataset is composed of 88 paired normal and tumor samples, where 21 samples are from normal tissues and the rest are obtained from tumor tissues, spanning 11 different tumor types. The mRNA expression was measured with microarray Hu6800 and Hu35KsubA GeneChips (Affymetrix, Santa Clara, CA) (Ramaswamy *et al.*, 2001), while the miRNA expression (GSE2564) was obtained using the flow cytometry technique (Lu *et al.*, 2005). The final compiled expression data used in GenMiR++ and TaLasso consist of 114 human miRNAs and 16 063 mRNAs.

The second dataset is a hepatocellular carcinoma (HCC) dataset downloaded from TCGA Data Portal (https://tcga-data.nci.nih.gov). It consists of matched miRNA and mRNA expression data from miRNA-seq and RNA-seq experiments, respectively. The information on patient survival outcome was also obtained and mapped based on the patient ID. The dataset includes 123 patients, among which 50 patients were alive at the time of last follow-up.

2.5 Methods evaluation

2.5.1 Experimentally validated targets

To evaluate our model, we downloaded experimentally validated miRNA targets from the miRTarBase as the gold standard. MiRTarBase 4.5 includes 14 659 experimentally verified miRNA–mRNA interactions between 358 miRNAs and 8400 target genes in human (Hsu *et al.*, 2014). These interactions have been validated by low-throughput assays including reporter assay and Western blot, or high-throughput transcriptomic experiments, such as microarray and pSILAC experiments. The sensitivity and specificity of the prediction method were systematically evaluated by the receiver operating characteristic (ROC) curve and the area under the curve (AUC), using the ROCR package in R (Sing *et al.*, 2005).

2.5.2 Gene ontology enrichment analysis

We extracted the Gene Ontology Biological Process (GO-BP) annotations from the Gene Ontology Annotation database. The GO terms with less than five genes or with no biological Data available were removed, giving totally 2321 GO-BP terms. We examined the GO categories enrichment for the predicted targets of each miRNA in each prediction method. The GO enrichment P-values were calculated by the hyper-geometric test, subject to the False Discovery Rate (FDR) control by Benjamini and Hochberg (BH) multiple testing procedure. The GO terms with FDR < 0.1 were considered as significantly enriched by predicted miRNA target sets.

2.5.3 Examining the clinical relevance of predicted targets

As an additional evaluation metric, we examined whether the predicted miRNA targets were biologically meaningful by analyzing the correlation between their gene expression profiles and the samples. Specifically, for the MCC dataset, we examined the association between the expression levels of each target gene with the tumor samples. The signal-to-noise (S2N) statistics was obtained for each gene, as in a previous study (Ramaswamy *et al.*, 2001). Here, the S2N of each gene was calculated as $(\mu_{tumor} - \mu_{normal})/(\sigma_{tumor} + \sigma_{normal})$, where μ and σ represent the mean and standard deviation of expression levels of each gene, respectively. The permutation test was then used to calculate whether the gene is a statistically significant tumor marker. 1000 permutations of the sample labels were performed on the dataset, and the S2N value was recalculated for each gene in each permutation. The gene is considered to be statistically significant if the observed S2N value exceeds the permuted ones at least 99% of the time ($P \leq 0.01$).

For the HCC dataset, since we have the patient clinical information available, we examined whether the targets were strongly associated with patient survival outcome. For each of the target, we divided the patients into two equal groups based on their expression values: the high-expression group and the low-expression group. We used the 'Survival' package in R to calculate the Kaplan-Meier survival curved. For each gene, its statistical significance was determined with the BH multiple testing procedure for the P-values obtained from log-rank test. The genes with FDR < 0.2 were considered as having significant association between their expression values and the liver cancer patient outcome.

2.5.4 Comparison with other methods

We first compared the performance of our adaptive Lasso (named AdaLasso) method with six other methods by evaluating their sensitivity and specificity according to the overlap between the predicted targets and the set of experimentally validated targets in miRTarBase. Among the six methods, GenMiR++ (Huang *et al.*, 2007), Lasso (Lu *et al.*, 2011) and TaLasso (Muniategui *et al.*, 2012) combine the sequence-based predictions and expression profiles, but neglect the full spectrum of sequence features. For the other three methods, the Context Score from TargetScan, the hybridization energy (ΔG), and the accessibility energy ($\Delta\Delta G$) are based on sequence information alone and do not consider expression data. Therefore, the miRNA–mRNA interactions predicted by each of the methods were ranked by their corresponding feature values, with the lowest value ranking on the top. According to the comparison results, we picked the best-performing sequence-based method, and then evaluated the targets predicted from this method and those from the expression-based methods by examining their clinical relevance (as described in detail in Section 2.5.3).

3 Results

3.1 Expression datasets

We applied the AdaLasso method on two datasets composed of miRNA and mRNA paired expression data. With a preprocessing step, the overlap between the MCC dataset and the predicted targets by TargetScanHuman 6.1 yielded totally 65 081 putative miRNA–mRNA pairs between 104 miRNAs and 9559 mRNAs (Muniategui *et al.*, 2012). Among them, 83 miRNAs with 1244 miRNA–mRNA pairs have biological verifications in the miRTarBase. For the HCC dataset, the intersection of the data with the predicted targets by TargetScanHuman yielded a final dataset with 207 515 putative miRNA–mRNA interactions between 170 miRNAs and 17 201 mRNAs. Among them, 102 miRNAs with 2470 miRNA–mRNA interactions have been experimentally validated, as listed in the miRTarBase.

3.2 Adaptive Lasso performs better on experimentally validated miRNA targets recovery

As a commonly applied penalty for variable selection, adaptive Lasso could perform differently depending on the choice of regularization parameter λ. In addition, the feature transformation will determine the magnitude of the feature values, which in turn regulate their corresponding Lasso penalty in the model. Therefore, the feature transformation is critical to the variable selection problem. We tuned the regularization parameter λ and the scale factor b simultaneously using the two-dimensional cross validation technique for each feature set individually. Specifically, we applied leave-one-out cross validation to test the parameter values. The optimal λ and b corresponding to the minimum square error (MSE) in testing samples were selected as tuning parameters. Based on the MSE results, the optimal parameter b was obtained as 6.5, 0.1 and 0.2 for the context score, ΔG and $\Delta\Delta G$ features, respectively. The difference of scale factors among the three features may be explained by the ranges of their corresponding feature values. We have found the values of ΔG and $\Delta\Delta G$ are scattered away from zero, and the values of context scores are more concentrated and closer to zero. Therefore, the context score feature utilizes a larger scale factor to transform its values to a range comparable to that of the other two feature sets.

Given the tuning parameters, we can use the Equation (5) to obtain values of β and w, where β quantifies the relationship between each pair of miRNA and mRNA, and ω indicates the weights of each feature. To assess the performance of our approach, we used experimentally validated target pairs from miRTarBase to examine the sensitivities and specificities for our model and other well-known methods. Correspondingly, we have calculated the ROC curves and the areas under ROC curve (AUC) for the optimal set of parameter (λ, b). In MCC dataset, our AdaLasso approach yielded an AUC score of 0.696 (Fig. 2A and Supplementary Table S1). For comparison, other methods that were based on sequence information only (Context Score, ΔG and $\Delta\Delta G$), and those combined expression information with sequence-based prediction (TaLasso, Lasso and GenMiR++) were also applied on the MCC dataset. Their corresponding ROC curves are shown in Figure 2A, and the AUC scores are listed in Supplementary Table S1. We found our AdaLasso approach that combined three sequences features and the miRNA/mRNA expression profiles had superior performance than the other methods according to the comparison of their sensitivity and specificity. In particular, TaLasso, GenMiR++, and Lasso yielded very low AUC scores of 0.537, 0.530 and 0.528, respectively. Note that the existing databases on miRNA target information are far from completeness. In addition, the miRNA regulation is dynamic and it is not clear to what extent the dynamic miRNA–mRNA interactions align with the information in existing databases, which may explain the low sensitivity and specificity obtained in these expression-based methods that do not fully utilize the sequence feature information. Although the Context Score method yielded a high AUC score (0.641), its performance was inferior to ours. These results indicate that combining the full spectrum of multiple types of sequence features with miRNA/mRNA expression data is critical in improving the model performance on identification of miRNA–mRNA interacting pairs. The same comparison was performed for the HCC dataset. We observed similar results as those for the MCC dataset, in that our AdaLasso method achieved the highest AUC score than both sequence-based methods and the expression-based ones (Fig. 2B and Supplementary Table S1). In addition, we note a Leukemia dataset (LDS), available in GEO (GSE14834) was used to evaluate the TaLasso approach. To obtain a more comprehensive comparison, we included this dataset here for further evaluating the performance of different methods on recovering experimentally validated miRNA targets. The results of using the LDS dataset were consistent with those obtained from the MCC and HCC datasets (Supplementary Fig. S1 and Table S1).

3.3 The contribution of sequence features for miRNA target prediction in adaptive Lasso

We further evaluated the effect of sequence features for predicting functional miRNA–mRNA interactions. We selected the putative

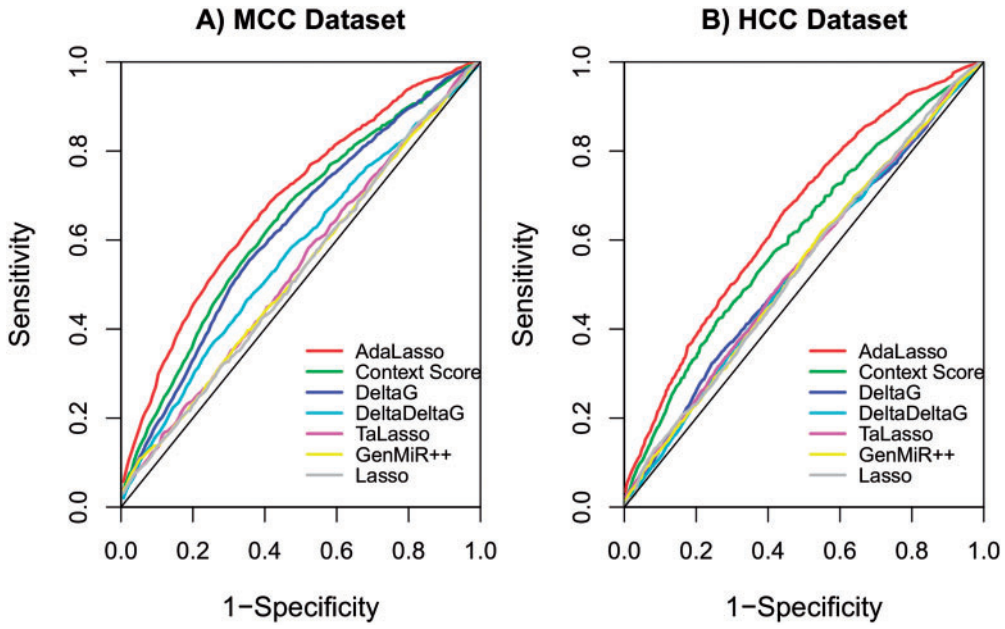

Fig. 2. ROC plots based on the experimentally-validated miRNA-mRNA pairs from the miRTarBase for the **(A)** MCC dataset; **(B)** HCC dataset. The diagonal line serves as a reference line

miRNA targets identified from each method according to their scores: one candidate group contains the top 1000 miRNA–mRNA interaction pairs with the highest prediction scores, and the other candidate group contains 1000 pairs with the lowest prediction scores. Their corresponding sequence features were compared based on the cumulative frequency analysis. As shown in the Figure 3 for the MCC dataset, the AdaLasso approach has the most differentiated feature values between the high- and the low-scoring groups. It is demonstrated the high-scoring miRNA-target interactions from AdaLasso always have lower feature values than those in the low-scoring candidate group, which is consistent with the well-known miRNA targeting rules (Kertesz *et al.*, 2007; Krek *et al.*, 2005). However, the results from TaLasso do not always follow this inversed relationship. In fact, some target pairs in the high-scoring group even have higher context scores than those in the low-scoring group, which to some degree violates the rationale of functional miRNA targeting. This may explain the relatively poor performance of TaLasso compared with our method. For the HCC dataset, the cumulative frequency plot analysis on high- and low-scoring interactions groups also showed the largest difference of their respective sequence feature values for AdaLasso compared with other methods (Supplementary Fig. S2). We also compared the contribution of each individual feature on our AdaLasso approach by examining the learned feature weights of ω. In the results, the three features (context score, ΔG, and $\Delta\Delta G$) had similar weights (0.36, 0.33 and 0.31, respectively). It agrees with the previous insights on miRNA target predictions that the thermodynamic stability, site accessibility and context score are all important for improving miRNA target prediction (Wang *et al.*, 2014; Xu *et al.*, 2014b). For the HCC dataset, the learned feature weights of ω are 0.35, 0.32 and 0.33 for context score, ΔG, and $\Delta\Delta G$, respectively, which are comparable to their values in the MCC data.

3.4 GO enrichment analysis

To further investigate the function of our predicted targets and the potential miRNA regulatory roles, we extracted the GO-BP annotations from the GO annotation, and examined the GO term enrichment for the miRNA target sets obtained by TaLasso, GenMiR++ and our approach, respectively. The GO enrichment *P*-values were calculated by the hyper-geometric test, followed by multiple hypotheses testing correction. For each miRNA and each prediction method, we examined the GO enrichment in the target gene sets predicted by the given method for the given miRNA. The GO terms with FDR < 0.1 were considered as significantly enriched. For the MCC dataset, GenMiR++ yielded totally 400 significantly enriched GO functional terms with FDR < 0.1, whereas TaLasso and our approach resulted in 440 and 468 significantly enriched GO terms, respectively. For the HCC dataset, GenMiR++, TaLasso and our approach obtained 451, 540 and 595 significantly enriched GO terms, respectively. The comparison results indicate that the targets predicted by our adaptive Lasso-based approach have more biologically meaningful annotations.

3.5 Clinical relevance of top predicted targets

We also examined whether there is a strong association between the gene expression profiles of targets and the samples. The results were summarized in Table 1. For MCC dataset, we only identified 7% of all genes that were differentially expressed between normal and tumor samples by using the S2N metric (see 'Methods' section). These genes were considered as statistically significant tumor

Table 1. The comparison of AdaLasso with other methods (Lasso, TaLasso, GenMiR++ and Context Score) in identifying clinically relevant targets

	MCC dataset				HCC dataset			
Methods	Sig.	Pred	%	p-value	Sig	Pred	%	p-value
AdaLasso	195	856	22.8	7.38E-35	119	899	13.2	6.30E-53
TaLasso	170	847	20.0	1.50E-24	98	858	11.4	4.64E-40
GenMiR++	128	845	15.1	3.42E-10	86	876	9.8	7.40E-26
Lasso	141	873	16.1	5.63E-13	83	959	8.7	3.12E-20
Context Score	124	859	14.4	1.06E-08	96	846	11.3	1.45E-35

Sig., the number of predicted targets whose expression profiles are associated with tumor samples (for MCC dataset) or patient survival outcome (for HCC dataset); Pred, the number of predicted targets; %, the percentage of significant targets in the prediction results.

P-values are obtained from Chi-squared test.

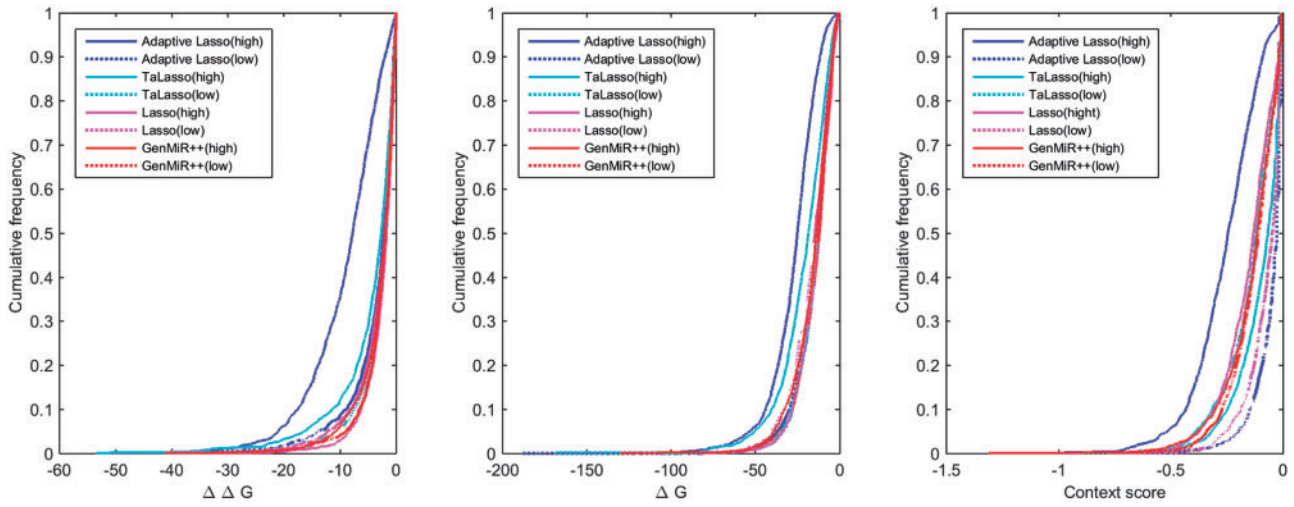

Fig. 3. The comparison of adaptive Lasso with other methods (Lasso, TaLasso and GenMiR++) for investigating the effect of sequence features in miRNA target prediction on the MCC dataset. The candidate miRNA-target interactions obtained from MCC are grouped according to their respective prediction scores, as the high scoring group (high) and the low scoring group (low). The cumulative frequency of each sequence feature is plotted. The x-axis represents the values for each sequence feature

markers. We then extracted the top 1000 miRNA–mRNA pairs predicted by our AdaLasso method. Among the 856 unique genes from these pairs, 22.8% of them can be considered as significant tumor markers (P = 7.38E-35, chi-squared test). As the Context Score method outperformed two other sequence-based methods (ΔG and $\Delta\Delta G$) in recovering experimentally validated targets, we also examined the top 1000 miRNA–mRNA pairs predicted by Context Score and found it yielded smaller percentage of significant tumor markers compared with all the expression-based methods for MCC dataset (Table 1). We note that the Context Score does not consider expression data, so it only represents static miRNA–mRNA interactions and may not capture the dynamic miRNA regulatory effects in a particular phenotype or condition. This may explain the small number of genes whose expression profiles are significantly associated with tumor samples by the Context Score method. For the HCC dataset, we also extracted the top 1000 interaction pairs predicted by each method, and performed the survival analysis on the unique genes from the top interactions. We found 13.2% of the top targets obtained from our AdaLasso method demonstrated significant association between their expression values and patient survival outcome. In contrast, only 11.4, 9.8, 8.7 and 11.3% of the top targets from TaLasso, GenMiR++, Lasso and Context Score can be considered as having the significant association, respectively (Table 1). As the control, we only found 2.9% of all genes included in the dataset demonstrated the significant association between their expression profiles and the liver cancer survival outcome. These results indicated that our AdaLasso approach outperformed others in effectively identifying clinically relevant targets.

4 Discussion and conclusion

In this article, we have proposed an adaptive form of Lasso for functional miRNA–mRNA interaction prediction, using the full spectrum of sequence features and the expression profiles of miRNAs and mRNAs. Our method is based on three assumptions: First, the number of predicted interactions is small. This assumption is ensured by the sparse solution of the linear regression problem. Second, we assume miRNAs down-regulate their corresponding mRNA targets. This assumption could lead to a potential problem that the targets affected by translation repression only would not be identified by our method. For such targets, the relevant miRNAs might not inhibit the mRNA expression profiles, but would reduce the protein levels (Selbach *et al.*, 2008; Zhuang *et al.*, 2015). It would be more comprehensive to examine the changes at both mRNA and protein levels if datasets are available. Third, the full spectrum of sequence features can make a significant contribution on improving target prediction. To validate this assumption, we have compared the feature values in high-scoring miRNA-target interactions with those in low-scoring ones. The results indicate the notion of a positive support from sequence features in adaptive Lasso procedure.

When compared with the existing computational methods that integrate expression data with sequence rule-based prediction, our method has demonstrated the utilities of three types of sequence feature information, including the context scores, the thermodynamic stability and the accessibility energy of the target sites. Our method based on the adaptive Lasso procedure automatically learns the relative contribution of each type of sequence features and dynamically integrates the features into expression profiles, while optimizing the expression support for predicting the miRNA–mRNA interactions. This offers a suitable solution for a challenging problem on how to integrate individual features for miRNA target prediction. Most current miRNA target prediction tools tend to consider only a single feature type, leading to different candidate sets of miRNA targets. According to the ROC analysis of predicted miRNA–mRNA interactions on the experimentally-validated targets, the GO analysis and the clinical relevance of predicted miRNA targets, our model outperforms other methods in real dataset applications. It has been well known that plant miRNAs often have targets with perfect or near-perfect sequence complementarity. In contrast, animal miRNA target sites only exhibit partial complementarity (Axtell *et al.*, 2011). The distinct pairing requirements and target breadth of animal and plant miRNAs suggest simpler target identification in plants than that in animals. Therefore, our method proposed here should not be used to predict miRNA–mRNA interactions in plants, but can be broadly applied for miRNA target predictions in other animal species.

Besides the three sequence features examined in our model, the degree of target site conservation can also be used as an additional sequence feature to represent the conservation of the target sites. However, the coverage of this feature in existing databases is significantly lower than the three features we used in this study. Therefore, to achieve high coverage of prediction results, we chose not to incorporate the target site conservation information in our methods. We also note that, with years of intensive biomedical research, a large amount of experimentally validated miRNA targets information has been accumulated in available databases. Therefore, the future work will focus on leveraging the prior known knowledge of miRNA targets to improve the prediction accuracy, since our adaptive form of regularized regression approach is capable of efficiently incorporating data from different sources into the model.

There has been increasing evidence that many miRNAs are generated as a family of related isomers that differ by a small number of bases at the 5′ and 3′ ends of the miRNA, termed isomiRs. Current sequence-based prediction programs only use single miRNA sequence, so they may underestimate the impact of miRNA isoforms on target regulation (Neilsen *et al.*, 2012). The expression of isomiRs varies between cell types and tissues, indicating their functional importance can be different among different cellular conditions (Tan *et al.*, 2014). With more expression profiles and the sequence information of isomiRs available, we expect our method is applicable for studying the miRNA–mRNA interactions at the isomiR levels as well.

Supplementary Data

Supplementary data are available at *Bioinformatics* online.

Funding

This work is supported in part by National Institutes of Health (R01 LM010022) and the seed grant from the University of Texas Health Science Center at Houston.

Conflict of Interest: none declared.

References

Axtell,M.J.*et al.* (2011) Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.*, **12**, 221.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297.

Betel,D. *et al.* (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**(Database issue), D149–D153.

Calin,G.A. *et al*. (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. USA*, **101**, 2999–3004.

Cao,S. *et al*. (2014) A Unified sparse representation for sequence variant identification for complex traits. *Genetic Epidemiol.*, **38**, 671–679.

Chen,D. *et al*. (2014) miR-100 induces epithelial-mesenchymal transition but suppresses tumorigenesis, migration and invasion. *PLoS Genetics* **10**, e1004177.

Fabian,M.R.*et al*. (2010) Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem,*. **79**, 351–379.

Friedman,R.C. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Garcia,D.M. *et al*. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.

Grimson,A. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Hsu,S.D. *et al*. (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**(Database issue), D78–D85.

Huang,J.C. *et al*. (2007) Bayesian inference of MicroRNA targets from sequence and expression data. *J. Comput. Biol.*, **14**, 550–563.

Huntzinger,E. and Izaurralde,E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat. Rev. Genet.*, **12**, 99–110.

Iorio,M.V. and Croce,C.M. (2012) MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol. Med.*,**4**, 143–159.

Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

Kim,S.J. *et al*. (2007) An interior-point method for large-scale l 1-regularized least squares. *IEEE J. Select Topics Signal Process*, **1**, 606–617.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**(Database issue), D152–D157.

Krek,A. *et al*. (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.

Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**(Web Server issue), W451–W454.

Lee,G. *et al*. (2011) Drosophila caspases involved in developmentally regulated programmed cell death of peptidergic neurons during early metamorphosis. *J. Comp. Neurol.*, **519**, 34–48.

Lee,S. *et al*. (2010). Adaptive multi-task Lasso: with application to eQTL detection. In: Lafferty, J.D. *et al*. (eds.). **Vol. 23**, *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Red Hood, NY, pp. 1306–1314.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Li,Y. *et al*. (2014) A probabilistic approach to explore human miRNA targetome by integrating miRNA-overexpression data and sequence information. *Bioinformatics*, **30**, 621–628.

Lu,J. *et al*. (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.

Lu,Y. *et al*. (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, **27**, 2406–2413.

Maragkakis,M. *et al*. (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.

Muniategui,A. *et al*. (2012) Quantification of miRNA-mRNA interactions. *PloS One*, **7**, e30766.

Neilsen,C.T. *et al*. (2012) IsomiRs–the overlooked repertoire in the dynamic microRNAome. *Trends Genet.*, **28**, 544–549.

Ramaswamy,S. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA*, **98**, 15149–15154.

Rehmsmeier,M. *et al*. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Selbach,M. *et al*. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Tan,G.C. *et al*. (2014) 5′ isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.*, **42**, 9424–9435.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B*, **58**, 267–288.

Wang,Z. *et al*. (2013) Incorporating prior knowledge into Gene Network Study. *Bioinformatics*, **29**, 2633–2640.

Wang,Z. *et al*. (2014) A Bayesian framework to improve microRNA target prediction by incorporating external information. *Cancer Inform.*, **13**(Suppl 7), 19.

Xu,W. *et al*. (2014a) identifying microRNA targets in different gene regions. *BMC Bioinformatics*, **15**(Suppl. 4), 11.

Xu,W. *et al*. (2014b) The Characterization of microRNA-mediated gene regulation as impacted by both target site location and seed match type. *PloS One*, **9**, e108260.

Zhuang,X. *et al*. (2015) Integrated miRNA and mRNA expression profiling to identify mRNA targets of dysregulated miRNAs in non-obstructive azoospermia. *Sci. Rep.*, **5**, 7922.