BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 17 2011, pages 2406–2413
doi:10.1093/bioinformatics/btr410

Systems biology

Advance Access publication July 8, 2011

A Lasso regression model for the construction of microRNA-target regulatory networks

Yiming Lu[1], Yang Zhou[1], Wubin Qu[1], Minghua Deng[2] and Chenggang Zhang[1,*]

[1]Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Beijing 100850 and [2]LMAM, School of Mathematical Sciences and Center for Theoretical Biology, Peking University, Beijing 100871, PR China

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs have recently emerged as a major class of regulatory molecules involved in a broad range of biological processes and complex diseases. Construction of miRNA-target regulatory networks can provide useful information for the study and diagnosis of complex diseases. Many sequence-based and evolutionary information-based methods have been developed to identify miRNA–mRNA targeting relationships. However, as the amount of available miRNA and gene expression data grows, a more statistical and systematic method combining sequence-based binding predictions and expression-based correlation data becomes necessary for the accurate identification of miRNA–mRNA pairs.

Results: We propose a Lasso regression model for the identification of miRNA–mRNA targeting relationships that combines sequence-based prediction information, miRNA co-regulation, RISC availability and miRNA/mRNA abundance data. By comparing this modelling approach with two other known methods applied to three different datasets, we found that the Lasso regression model has considerable advantages in both sensitivity and specificity. The regression coefficients in the model can be used to determine the true regulatory efficacies in tissues and was demonstrated using the miRNA target site type data. Finally, by constructing the miRNA regulatory networks in two stages of prostate cancer (PCa), we found the several significant miRNA-hubbed network modules associated with PCa metastasis. In conclusion, the Lasso regression model is a robust and informative tool for constructing the miRNA regulatory networks for diagnosis and treatment of complex diseases.

Availability: The R program for predicting miRNA–mRNA targeting relationships using the Lasso regression model is freely available, along with the described datasets and resulting regulatory network, at http://biocompute.bmi.ac.cn/CZlab/alarmnet/. The source code is open for modification and application to other miRNA/mRNA expression datasets.

Contact: zhangcg@bmi.ac.cn

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

Received on April 5, 2011; revised on June 15, 2011; accepted on July 3, 2011

*To whom correspondence should be addressed.

1 INTRODUCTION

MicroRNAs (miRNAs) are 20–24 nt RNAs that derive from distinctive hairpin precursors in animals, plants and fungi, and can bind to complementary sequences on target mRNAs. In recent years, miRNAs have emerged as a major class of regulatory elements of gene expression. Computational predictions of miRNA targets estimate that each miRNA regulates hundreds of different mRNAs, and >60% of human protein-coding genes are subject to miRNA regulation (Friedman *et al.*, 2009). MiRNAs are involved in a broad range of biological processes, including development, differentiation, apoptosis and proliferation (Bartel, 2004; Harfe, 2005). Several studies have shown that miRNAs are involved in the initiation and progression of cancer (Lu *et al.*, 2005; Volinia *et al.*, 2006). Significantly differing miRNA profiles in various types of tumors suggests that miRNA profiling has diagnostic and perhaps prognostic potential (Garzon *et al.*, 2009).

MiRNAs use base pairing to guide RNA-induced silencing complexes (RISCs) to specific mRNAs with imperfect complementary sequences. The recruitment of RISC commonly leads to the repression of targeted mRNA through inhibition of translation initiation and/or miRNA-mediated mRNA destabilization (Bushati and Cohen, 2007; Eulalio *et al.*, 2008). The Argonaute protein in the RISC complex is crucial for gene silencing to occur for its capability of endonucleolytic cleavage. In human, there are four Ago proteins, where Ago2 is the sole protein required for the RISC activity (Liu *et al.*, 2004; Meister *et al.*, 2004). A miRNA can directly destruct the target mRNA through mRNA cleavage when the pairing is extensive (Liu *et al.*, 2004). This kind of repression dominates in plants (Jones-Rhoades *et al.*, 2006). In animals, however, it was once thought miRNAs repress protein output with little or no influence on mRNA levels, because in animals only a few targets satisfy the extensive pairing required for cleavage (Olsen and Ambros, 1999; Wightman *et al.*, 1993). However a number of microarray data showed that miRNAs decrease the levels of many targeted mRNAs (Giraldez *et al.*, 2006; Krutzfeldt *et al.*, 2005; Lim *et al.*, 2005). Only recently it became clear that, similar as in plants, miRNAs predominantly exert their effects in mammalian cells through decreasing the levels of target mRNAs, and the mRNA destabilisation that accounts for most (≥84%) of the decreased protein production (Guo *et al.*, 2010; Swami, 2010).

A number of computational prediction methods have been introduced to identify the targets of miRNAs. The first generation of miRNA target prediction methods were primarily based on Watson–Crick base pairing rules and free energy calculations for

the miRNA/mRNA duplex, with very limited incorporation of experimental data. The resulting predictions differed widely between methods. Current miRNA target predictions methods are based primarily on systematic target site mutation experiments and adopted the 'seed pairing' approach, which could improve the accuracies of the predictions. However, they are still acknowledged to have issues with specificity, and the convergences of the predictions are also unsatisfactory (Rajewsky, 2006; Sethupathy *et al.*, 2006b).

It has been suggested that integrating the miRNA and mRNA expression data into the prediction could play an important role in the identification of miRNA targets. A natural thought of using the miRNA/mRNA expression data is to capture the negative correlation relationship between a miRNA and its target mRNA, based on which several methods have been developed recently (Ritchie *et al.*, 2010; Volinia *et al.*, 2010). This marginal approach is intuitive and simple but could identify very few known target pairs (Stanhope *et al.*, 2009). An improvement of the marginal approach is developed using AIC-optimal regression model and Ago expression information (Stanhope *et al.*, 2009). However, like the marginal approach, the AIC-optimal model investigates the miRNA–mRNA relationship individually, failing to consider the effects of multiple miRNAs targeting and their competition in available binding sites (Doench and Sharp, 2004; Krek *et al.*, 2005; Rajewsky, 2006).

As with transcriptional regulation, an mRNA can be simultaneously suppressed by more than one miRNA (Rajewsky, 2006). Both *in vitro* experiments (Doench and Sharp, 2004) and the observation that many 3′-untranslated regions (3′-UTRs) had predicted targets for different miRNAs demonstrate that a number of miRNAs act together to regulate a target mRNA (Doench and Sharp, 2004; John *et al.*, 2004; Krek *et al.*, 2005). Besides, these miRNAs can repress target mRNA to varying orders of magnitude (Grimson *et al.*, 2007; Lim *et al.*, 2005). Features like target site types (Bartel, 2009), target mRNAs abundance (Arvey *et al.*, 2010) and AU-rich nucleotide composition near the site (Grimson *et al.*, 2007) can affect the efficacies of miRNA repression. Hence, reliable identification of miRNA–target relationship requires consideration of different layers of information, including sequence-based information (Krek *et al.*, 2005; Lewis *et al.*, 2005), co-regulation of miRNAs, RISC availability (Stanhope *et al.*, 2009) and miRNA/mRNA abundance (Gennarino *et al.*, 2009; Rajewsky, 2006).

In this article, we develop a linear regression model to investigate one mRNA simultaneously regulated by multiple targeting miRNAs using miRNA/mRNA expression data. To control variable number and improve model specificity, the sequence-based predictions are used to determine the initial variables of the model. In particular, we use a Lasso regression algorithm (Tibshirani, 1996) that imposes a constraint on the coefficients, by which the coefficients of false miRNA–mRNA pairs will be set to zero. The central idea of the model is consistent with the nature of different miRNAs coordinately regulating a target mRNA and their potential competition in binding sites (Doench and Sharp, 2004; Krek *et al.*, 2005; Rajewsky, 2006). We assess the performance of the Lasso regression approach by comparing it with two related identification methods which also use miRNA/mRNA expression data, the marginal miRNA–mRNA comparison method and the AIC-optimal regression method (Stanhope *et al.*, 2009). We find that the Lasso regression model has the following features: (i) this model can recover a larger number of known miRNA–target pairs, especially when compared with the marginal comparison method, while also achieving a significant improvement in identification specificity; (ii) the constraint imposed on coefficients enables the coefficients to reflect the true magnitudes of miRNAs regulation in tissues; (iii) unlike the marginal method and AIC-optimal regression method, the Lasso regression approach can simultaneously evaluate all possible targeting miRNAs of an mRNA in one model, making it a rapid and robust tool to construct the miRNA-target regulatory networks, especially when dealing with cross-platform and high-throughput expression data.

2 METHODS

2.1 Regression modelling

2.1.1 Sequence-based computational prediction First, sequence-based computational prediction was used to determine the initial variables (miRNAs) in the regression model. Previous comparison of several miRNA target prediction algorithms showed that the overlap between the predictions of TargetScan and PicTar are significantly higher than with any other algorithms (roughly 80–90% identical sites when run on the same dataset of 3′-UTRs) (Rajewsky, 2006). In addition, by comparing the predictions with ∼130 experimentally confirmed miRNA–mRNA regulatory relationships in *Drosophila melanogaster*, PicTar was found to have higher accuracy and sensitivity than other algorithms. Therefore, in this study, we use TargetScan (Grimson *et al.*, 2007; Lewis *et al.*, 2005) and PicTar (Sethupathy *et al.*, 2006a) to perform the computational prediction. Only those miRNAs simultaneously predicted as targets by these two algorithms are selected to be the initial miRNAs in the model.

Thus, let y^i index the expression level of the i-th mRNA, j index the j-th Ago protein, x_k^i represent the expression level of the k-th potential targeting miRNA of the i-th mRNA simultaneously predicted by TargetScan and PicTar algorithms, P_i represent the number of all potential targeting miRNAs of the i-th mRNA and ε^i be a random error term assumed to be normally distributed. The original linear regression model can be written as follows:

$$y^i = \beta_0^i + \sum_{k=1}^{P^i} \beta_k^i x_k^i + \varepsilon^i \tag{1}$$

2.1.2 Evaluation of Ago expression terms It has been previously shown that combining RISC and miRNA expression data could reveal more negative relationships between targeting miRNAs and mRNAs (Stanhope *et al.*, 2009). As binding to RISC is essential for the miRNAs to repress target mRNAs and miRNAs tend to bind to the same target sites of RISC (Meister *et al.*, 2004; Schmitter *et al.*, 2006), it is reasonable to assume that concentration changes in the Ago 1, 3, 4 proteins will affect the ability of Ago2 to bind RISC and cleavage target mRNAs. Therefore, we consider each Ago protein and its potential effects on the concentration of target mRNA. Here, we assume that the mRNA levels of the Ago genes are positively related to their protein concentrations. Let Ago_j represent the j-th Ago mRNA level, and the regression model with Ago terms can be written as follows:

$$y^i = \beta_0^i + \sum_{j=1}^{4} \sum_{k=1}^{P^i} \beta_{jk}^i \mathrm{Ago}_j x_k^i + \varepsilon^i \tag{2}$$

Because the Ago 1, 3, 4 proteins are all regarded as Ago2's competitors in binding with RISC, we can merge the expressions of the three proteins into one term, Ago134. This transformation can reduce the number of coefficients and transform model (2) to the following:

$$y^i = \beta_0^i + \sum_{k=1}^{P^i} \beta_{2,k}^i \mathrm{Ago}_{2,k} x_k^i + \sum_{k=1}^{P^i} \beta_{134,k}^i \mathrm{Ago}_{134,k} x_k^i + \varepsilon^i \tag{3}$$

To further evaluate the effects of Ago proteins on the regression model, we will apply models (1) and (3) to the same datasets to study the contribution of Ago expression terms to the model performance.

2.2 Variable selection

By analysing the miRNA–mRNA regulatory relationships predicted by TargetScan and PicTar, we found that, on average, one mRNA could be potentially targeted by >30 miRNAs. Given that a considerable portion of the miRNA–mRNA relationships predicted by TargetScan and PicTar may be false positives, selection of initial variables is extremely important for the regression model. The Lasso algorithm is adopted to perform the variable selection procedure by estimating linear regression coefficients through L_1-constrained least squares. Specifically, it minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the existence of this constraint, the Lasso regression tends to produce some coefficients that are exactly 0 and hence can give robust and interpretable models. Specifically, for regression model (1), the corresponding constrained L_1 norm can be given by the following inequality:

$$\sum_{k=1}^{P^i} \left|\beta_k^i\right| < t \tag{4}$$

For regression model (2), the corresponding constrained L_1 norm is:

$$\sum_{k=1}^{P^i} \left(\left|\beta_{2,k}^i\right| + \left|\beta_{134,k}^i\right|\right) < t \tag{5}$$

Thus, if the variables do not indicate the real regulatory relationships of miRNAs and mRNA, the corresponding coefficients would converge to zero. This problem is equivalent to minimizing the following loss function in a typical *Lagrangian form* for model (1):

$$\hat{\beta}^i = \frac{1}{2}\sum_{l=1}^{N}\left(y^i - \beta_0^i - \sum_{k=1}^{P^i}\beta_k^i x_{k,l}^i\right)^2 + \lambda\sum_{k=1}^{P^i}\left(\beta_k^i\right) \tag{6}$$

and for model (3):

$$\hat{\beta}^i = \frac{1}{2}\sum_{l=1}^{N}\left(y^i - \beta_0^i - \sum_{k=1}^{P^i}\beta_{2,k}^i \mathrm{Ago}_{2,k,l} x_{k,l}^i - \sum_{k=1}^{P^i}\beta_{134,k}^i \mathrm{Ago}_{134,k,l} x_{k,l}^i\right)^2 + \lambda\sum_{k=1}^{P^i}\left(\beta_{2,k}^i + \beta_{134,k}^i\right) \tag{7}$$

Here $\lambda \geq 0$ is a complexity tuning parameter that controls the amount of shrinkage. The larger the value of λ is, the greater the amount of shrinkage. Hence, λ should be adaptively chosen to provide an estimate of expected prediction error.

Computationally, the Lasso model represents a quadratic programming problem, which can be tackled by standard numerical analysis algorithms. In practice, we use the R programming language (Ihaka and Gentleman, 1996; Team, 2008) and the 'lars' package (Efron *et al.*, 2004) to implement the Lasso regression and variable selection procedure. The basic procedure of the Least Angle Regression (LARS) algorithms in the 'lars' package is starting with all coefficients equal to zero and adding variable that currently most correlated with y in a specific direction one-by-one until all variables are in the model, as λ is varied from infinity to 0. To determine which subset of variables should be chosen from the collection of possibilities, a C_p-type statistic is derived to estimate the prediction error. Let $\hat{\mu}_k = X\hat{\beta}_k$ represent the k-step estimator of y, $\bar{\sigma}^2$ represent the ordinary least squares (OLS) estimate of variance and n represent the sample size. The C_p statistic can be given as follows:

$$C_p(\hat{\mu}_k) \doteq \left\|y - \hat{\mu}_k\right\|^2/\bar{\sigma}^2 - n + 2k \tag{8}$$

The LARS algorithm then determines which step minimizes C_p. Because miRNAs are known as negative regulators of target genes, only variables (miRNAs) with negative $\beta_{2,k}^i$ are selected as true regulatory miRNAs.

2.3 Model refinement

In practice, the sample size is usually small; hence, a subset of the true targeting miRNAs cannot be selected by the model in only one run of variable selection. To address this issue, we introduce a complementary improvement by adopting a multi-run Lasso regression procedure and ranking the selected miRNAs by their coefficients. First, we create two groups (I and II) and put all the initial variables in Group II. Then for each run, the Lasso procedure is performed as described above with the variables in Group II, and variables assigned non-zero coefficients are moved from Group II to Group I. This loop stops if all variables in Group I have zero coefficients or if Group I is empty. Next, the negative coefficients of selected variables are normalized, and the miRNAs are ranked according to their coefficients. As the selected coefficients are always negative, stronger regulatory relationships always correspond to greater absolute values and lower ranks. We define the rank score of the i-th miRNA based on the rank of its normalized coefficient in the set:

$$\mathrm{RankScore}^i = \frac{\mathrm{Rank\ of\ miRNA}^i}{\mathrm{Number\ of\ negative\text{-}coefficient\ miRNAs}} \times 100 \tag{9}$$

When the sample size is especially small (lesser than or equal to variables number), the Cp statistic cannot be efficiently calculated. In this situation, the multi-run Lasso regression procedure is slightly different to that described above. Instead of selecting non-zero coefficient variables in the step that minimizes Cp, in each run we only select the variable first entering the model (most correlated with y), and remove it from Group II, until the number of variables in Group II is sufficiently small to calculate Cp. Afterwards, the runs are the same as described above.

2.4 Expression data

Regression model (2) and the variable selection procedure were implemented using data from three studies in which both human mRNA and miRNA expression levels were measured on a set of tissue samples. We used two studies as the same as that used by the AIC-optimal regression method for the convenience of methods comparison. The first study of nasopharyngeal cancer (NPC) by researchers in Madison, WI, USA and elsewhere (Sengupta *et al.*, 2006, 2008) derived mRNA and miRNA profiling from 31 NPC and 10 normal tissue samples using the whole genome Affymetrix hgu133plus2 microarrays and custom cDNA arrays. The second study of miRNA expression patterns over a wide variety of tumor and normal tissue types by the Broad Institute (Lu *et al.*, 2005) derived data from 67 tissue samples of 10 different normal and tumor tissue types. Besides, we used the third study of integrative genomic profiling of human PCa conducted by Memorial Sloan-Kettering Cancer Center (MSKCC) (Taylor *et al.*, 2010) derived mRNA and miRNA profiling from 28 normal, 98 primary cancer and 13 metastatic cancer tissues samples using the Affymetrix Human Exon 1.0 ST Array and Agilent Human miRNA Microarray 2.0. All the above expression data were carefully parsed and formatted for the analysis of the Lasso regression model.

To identify the significant mRNAs between each group of the MSKCC dataset and measure the fold changes of miRNAs and mRNAs, we employed the Significance Analysis of Microarrays (SAM) algorithm (Tusher *et al.*, 2001) to statistically analyse the expression data. For each analysis, the SAM algorithm calculates the false discovery rates (FDRs) and q-value to determine the significance of genes. Details of the SAM analysis results are provided in Supplementary Tables S1–S4.

2.5 Known target pairs

To evaluate the Lasso regression model, we used two sets of known target pairs. One set was collected by Stanhope *et al*. from the TarBase miRNA target database (Sethupathy *et al.*, 2006a) and has been used to validate the AIC-optimal regression model. All of them had been previously validated through the use of mRNA- and protein-specific techniques such as polymerase chain reaction (PCR), luciferase reporters and immunoblotting and were represented in the Madison and Broad datasets (relationships that were only supported by microarray data were

not included). There were 76 known miRNA–mRNA pairs from the Broad dataset and 99 miRNA–mRNA pairs (including the previously reported 76 pairs) from the Madison dataset. We collected the second set of known miRNA–mRNA pairs by searching the literature for all experimentally validated miRNA–mRNA targeting relationships that have been demonstrated with PCa. In total, we collected 121 such pairs, and 106 of them were present in the MSKCC dataset. We used this set of known miRNA–mRNA pairs for measurement of the MSKCC dataset. All the miRNA–mRNA pairs are listed with the corresponding literature sources in Supplementary Table S5.

2.6 Randomization controls

To evaluate the specificity of the Lasso regression method, we applied this method repeatedly to a set of randomized control data and recorded the identification number for each repetition. Because the randomized control miRNA–mRNA pairs cannot guarantee fully excluding of targeting relationships even when known target pairs are rejected from those (it is possible that some miRNA–mRNA pairs are targeting but failed to be predicted by current methods), we constructed the randomized control data by randomly choosing one miRNA from the initial targeting miRNAs set of each regression model and resampling the observed miRNA levels in different tissue samples. The original and resampled miRNA data were then simultaneously put back to the initial miRNAs set that is to be applied to the Lasso regression method. For each dataset, we repeated the randomization procedure for 100 times recording the positive identification numbers and the corresponding rank scores. These distributions of rank scores of randomly sampled and validated miRNAs were used to generate the receiver operating characteristic (ROC) curves to examine the performance of rank score in classifying resampled miRNAs and real targeting miRNAs.

2.7 Network module enrichment evaluation

To evaluate the significance of the miRNA-core modules, we constructed 10 000 random networks according to each real network recording the occurrences and the size (number of nodes) of the miRNA-core modules. Each random network has the identical number of miRNAs, mRNAs and the relationships with the corresponding real network. The times that the occurrence of a certain size of module in random networks is not less than that of the real network were used to calculate the *P*-value of one module in a real network.

3 RESULTS

3.1 Lasso regression model performs better in miRNA-target identification

To evaluate the effect of the Ago expression terms in the models, we first assessed the performances of regression models (1) and (3) in identification of miRNA–mRNA targeting relationships. We used known target pairs and randomized control data to examine the sensitivities and specificities of these two models in three real miRNA/mRNA expression datasets as described in Section 2.4. We repeatedly applied the two models to each dataset for 100 times, and the ROC curves were plotted. As shown in Figure 1, in all three datasets, model (3) performs better than model (1) both on sensitivity and specificity when using 60 as a cut-off. The average areas under the ROC curves (AUCs) of model (3) for the Broad, Madison and MSKCC datasets were 0.75, 0.80 and 0.86, larger than the AUCs of model (1), which were 0.72, 0.70 and 0.78, respectively. This result demonstrates that combining Ago expression with miRNA/mRNA expression data is helpful in improving the model performance on identification of miRNA–mRNA targeting pairs.

We next compared the performance of the Lasso regression model with two other methods of the same type: the marginal comparison method and the AIC-optimal regression model, each of which also use miRNA/mRNA expression data to identify miRNA–mRNA pairs. The Lasso regression model integrates sequence-based prediction information, miRNA co-regulation and RISC availability data; thus, we anticipated it would be able to identify more known miRNA–mRNA pairs. We assessed the performances of these three methods using three expression datasets. As shown in Table 1, of the three datasets, the Lasso regression method (full set of selection) identified a significantly larger portion of known targeting pairs than the other two methods. The mean true detection rate (TDR) of the Lasso method for the three datasets was 43.7%, which is significantly higher than the mean TDRs for the marginal method (14.4%) and the AIC-optimal regression method (26.8%). When only considering the miRNAs whose rank scores were $\geq$60, a mean TDR of 32.3% across the three datasets is obtained, which is still significantly higher than

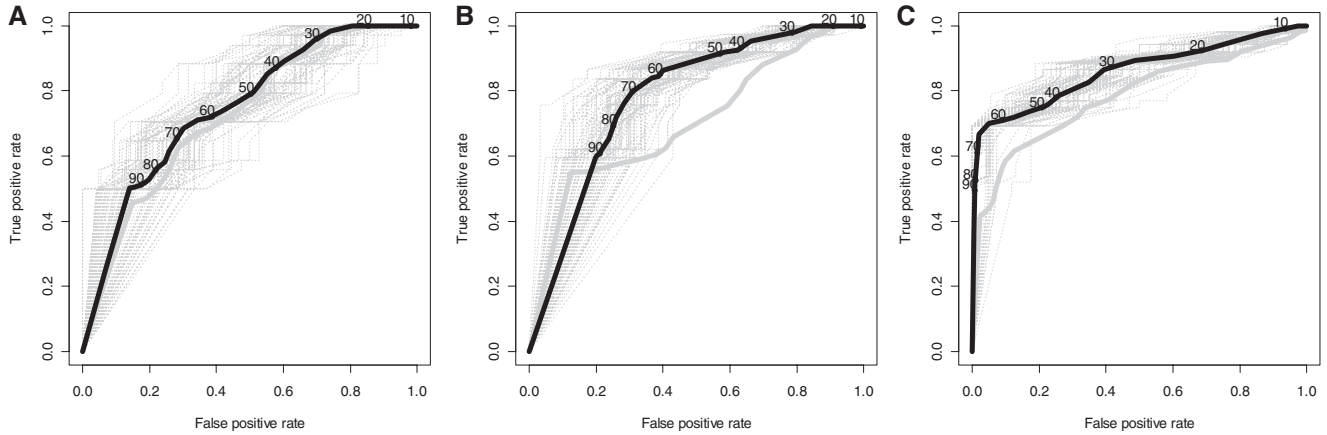

Fig. 1. ROC plots of the Lasso regression model in three datasets: Broad dataset (**A**), Madison dataset (**B**) and MSKCC dataset (**C**). The gray dash lines represent the ROC curves of 100 repeats, the black solid line represents the average ROC curve of model (3), and the gray solid line represents the average ROC curve of model (1). The numbers on the curves are different cut-offs of rank score. All the ROC curves are generated using the ROCR package (Sing *et al.*, 2005) in R.

Table 1. Performance comparisons among three different methods

Dataset	Sample size	Marginal comparison (%)	AIC-optimal model (%)	Lasso regression model (%)	Lasso regression model (Rank score ≥60) (%)
Broad	77	6/76 (7.9)	20/76 (26.3)	34/76 (44.7)	24/76 (31.6)
Madison	41	5/99 (5.1)	33/99 (33.3)	37/99 (37.4)	31/99 (31.3)
MSKCC	139	32/106 (30.2)	22/106 (20.8)	52/106 (49.1)	36/106 (34.0)

the TDR of the marginal method and slightly higher than that of the AIC-optimal regression method. This result further demonstrates that the Lasso regression model has higher sensitivity in detecting miRNA–mRNA targeting pairs.

We then examined the specificities of these models. Because the Lasso regression model employs L_1 coefficient constraints to select variables, we anticipated that it would show higher specificity. We found that, on average, ∼61.0% of non-targeting controls are filtered out by the Lasso regression variable selection procedure through directly setting their coefficients as 0. The ROC plots (Fig. 1) show that the rank score could well distinguish true targeting pairs from the remaining non-zero coefficient control pairs. By accepting the miRNAs whose rank scores were ≥60, the global specificities of the Lasso regression model for rejecting non-targeting control pairs in three datasets were ∼86, 86 and 97%. We note that the specificity of the Lasso regression model is outstanding among the three methods; the specificities of the AIC-optimal method, the best previous method of the same type and the marginal method were never >80% (Stanhope *et al.*, 2009) for the same datasets. Thus, it is clear that the Lasso regression model achieves significant improvements in both sensitivity and specificity for miRNA-target identification.

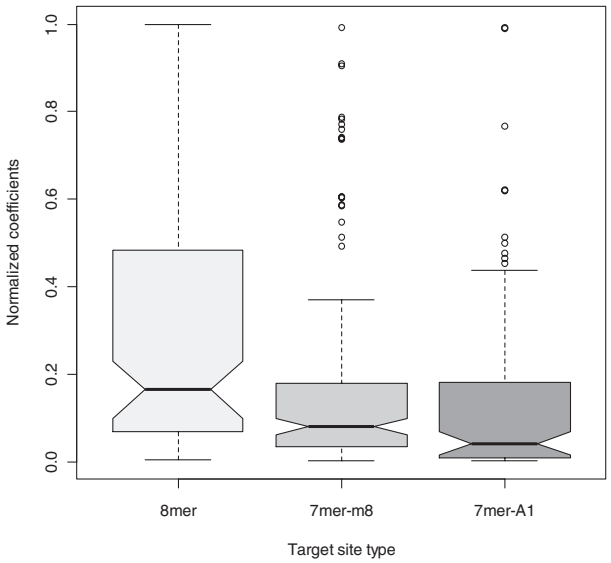

Fig. 2. Boxplot of the normalized coefficient distribution among three different miRNA–mRNA target types.

3.2 Constrained coefficients in the Lasso model reflect regulatory efficacies

The Lasso regression model combines multiple potential competing miRNAs in one regression model and imposes a constraint on their coefficients, which can provide more interpretable regression models by filtering out false positives and making the values of coefficients more biologically meaningful. We hypothesize that the constrained coefficients of miRNAs in the Lasso model may reflect their regulatory efficacies in the tissues studied. We tested this hypothesis by classifying the predicted miRNAs based on their target site types, which have been shown to have different levels of regulatory efficacy (Grimson *et al.*, 2007; Nielsen *et al.*, 2007). We considered three major types of miRNA–mRNA target sites: *8mer*, *7mer-m8* and *7mer-A1* sites. To conduct reliable verification and control noise, we used the Lasso regression model to predict the targeting miRNAs of the top 100 significant mRNAs from the MSKCC dataset, which has the largest sample size. All the predicted miRNA–mRNA pairs were divided into three groups based on their binding site types: the *8mer*, *7mer-m8* and *7mer-A1* groups, which were obtained from the TargetScan database. Overall, we obtained 99 *8mer* site pairs, 163 *7mer-m8* site pairs and 102 *7mer-A1* site pairs. Figure 2 shows the notched boxplot of the normalized coefficients distributions in three groups. We found that the mean coefficients of *8mer* site miRNAs were significantly higher than for two *7mer* sites, *7mer-m8* ($P = 2.531\text{e-}07$) and *7mer-A1* ($P = 6.191\text{e-}05$), and the coefficients of the *7mer-m8* site were also significantly higher than those of the *7mer-A1* site ($P = 0.01414$). The statistical significances were calculated using the Mann–Whitney U test. These observations were completely consistent with the characteristics of miRNA binding efficacy described by other groups using array data from miRNA transfection and miRNA/Dicer knockout experiments (Grimson *et al.*, 2007; Nielsen *et al.*, 2007). This result further supports our hypothesis that the coefficients of the Lasso regression model can be used to estimate the regulatory efficacy of miRNA–mRNA targeting pairs.

3.3 Constructing the miRNA regulatory network using the Lasso regression model

Because of the outstanding performance of the Lasso regression model in identifying miRNA–mRNA targeting relationships and reflecting real binding efficacies of miRNAs in tissue cells, we anticipated that the Lasso regression model could be a promising tool for constructing miRNA regulatory networks. Therefore, we constructed miRNA regulatory networks associated with PCa by analysing the MSKCC dataset with the Lasso model. The tissue samples of the MSKCC dataset were classified into three groups based on their clinical information: normal (28 samples), primaries (98 samples) and metastases (13 samples). Two miRNA regulatory networks were constructed by comparing the normal and primaries groups (also called Stage I of PCa) and the primaries and metastases groups (Stage II of PCa). To make the networks more specific, we

Table 2. Enriched modules in the PCa-related miRNA regulatory networks

Size	Occurrence	*P*-value	Hub miRNA	Fold change	KEGG pathway	Reference
Normal versus Primaries						
10	2	0.0097	hsa-miR-27b	−1.587	Wnt signalling pathway, Pathways in cancer	
			hsa-miR-200c	1.801	–	Schmalhofer *et al.* (2009)
9	3	0.022	hsa-miR-27a	−1.326	Vascular smooth muscle contraction	Sun *et al.* (2010)
			hsa-let-7b	1.174	–	Fu *et al.* (2011)
			hsa-miR-30b	1.243	ARVC, HCM, dilated cardiomyopathy	
8	5	0.0002	hsa-miR-29a	−1.108	ECM–receptor interaction, pathways in cancer, small cell lung cancer	Gebeshuber *et al.* (2009)
			hsa-miR-29b	−1.045	ECM–receptor interaction, pathways in cancer, small cell lung cancer	Steele *et al.* (2010)
			hsa-miR-32	2.145	–	
			hsa-miR-98	1.121	Wnt signalling pathway	
			hsa-miR-19a	1.632	ARVC, focal adhesion	
Primaries versus Metastases						
13	1	0.0011	hsa-miR-19b	−1.466	ARVC	
11	2	0.0003	hsa-miR-29b	−2.126	Focal adhesion, pathways in cancer, regulation of actin cytoskeleton	Steele *et al.* (2010)
			hsa-miR-19a	−1.294	ARVC, focal adhesion	
10	2	0.0052	hsa-miR-181a	1.692	–	Wang *et al.* (2011)
			hsa-miR-106b	−1.034	Axon guidance, focal adhesion	Slaby *et al.* (2010)
9	2	0.0673	hsa-miR-30c	−2.018	Pathways in cancer	Heinzelmann *et al.* (2011)
			hsa-miR-15b	1.347	MAPK signalling pathway	Loayza-Puch *et al.* (2010)
8	5	0.0071	hsa-miR-16	−1.725	Pathways in cancer	Takeshita *et al.* (2010)
			hsa-miR-27b	−3.477	Pathways in cancer	
			hsa-miR-141	1.127	Axon guidance, pathways in cancer	Gregory *et al.* (2008)
			hsa-miR-98	−1.736	Focal adhesion, MAPK signalling pathway	
			hsa-miR-30d	1.690	ARVC, HCM, dilated cardiomyopathy	Yao *et al.* (2010)

The significant KEGG pathways are annotated using DAVID software (Huang da *et al.*, 2009).

only considered the significant mRNAs (fold change $\geq$1.5, FDR $\leq$ 1.0e-06 and $q \leq$ 1.0e-06) for each pair of groups. These significant mRNAs were then processed by the Lasso regression procedure to identify the miRNA–mRNA targeting relationships, and only relationships with rank scores $\geq$60 were chosen to construct the miRNA regulatory networks. The full networks are available in the Supplementary Material.

We found that both networks were enriched with network modules taking miRNA as a regulatory hub when compared with randomized networks and these modules are involved in different biological pathways (Table 2). Figure 3 shows the subnetworks of the miRNA regulatory networks for two stages of PCa. In the subnetworks, we only show the most significant miRNA-hubbed network modules. Interestingly, comparison of the enriched modules in two stages of PCa revealed large differences, with more modules unique to each network than shared between them. For example, 6 of the 10 modules identified in Stage I (Normal versus Primaries) were not identified in Stage II (Primaries versus Metastases), while 7 of 11 modules in the Stage II network were not identified in Stage I network. In addition, the four common modules shared by the two networks also vary greatly in their targeted genes (as shown in Fig. 2). This reflects the great variety of roles that hub miRNAs play in the two stages of PCa. A recent study (Vallejo *et al.*, 2011) has even demonstrated that hsa-miR-200c, one of the unique hub miRNAs in the Stage I network, could inhibit PCa metastasis by targeting several genes such as ZEB1 that serve as biomarkers for PCa metastasis (Schmalhofer *et al.*, 2009). The significant overexpression of hsa-miR-200c in the Stage I network reveals its important role in the anti-metastasis process of Stage I PCa. Another hub miRNA, hsa-miR-29b, identified in both stages of PCa, was also reported to be associated with PCa metastasis by inhibiting the expression of MMP-2 and COL1A1. Additionally, many other hub miRNAs identified in these two networks have also been reported to be associated with cancer metastasis (Table 2) (Fu *et al.*, 2011; Gebeshuber *et al.*, 2009; Gregory *et al.*, 2008; Heinzelmann *et al.*, 2011; Loayza-Puch *et al.*, 2010; Slaby *et al.*, 2010; Steele *et al.*, 2010; Sun *et al.*, 2010; Takeshita *et al.*, 2010; Wang *et al.*, 2011; Yao *et al.*, 2010).

4 DISCUSSION AND CONCLUSION

In this study, we proposed a Lasso regression model to identify miRNA–mRNA targeting relationships, and we constructed miRNA regulatory networks representing two stages of PCa progression. By comparing our method with two other miRNA-target identification methods using three different datasets, we found that the Lasso regression model has considerable advantages in terms of both sensitivity and specificity. We also note that the regression coefficients in this model can be used to reflect the true regulatory efficacies in living cells, and we verified this hypothesis using the miRNA target site type data. By constructing the miRNA regulatory networks for two stages of PCa, we found that several significant miRNA-hubbed network modules are associated with PCa metastasis. In conclusion, the Lasso regression model is a

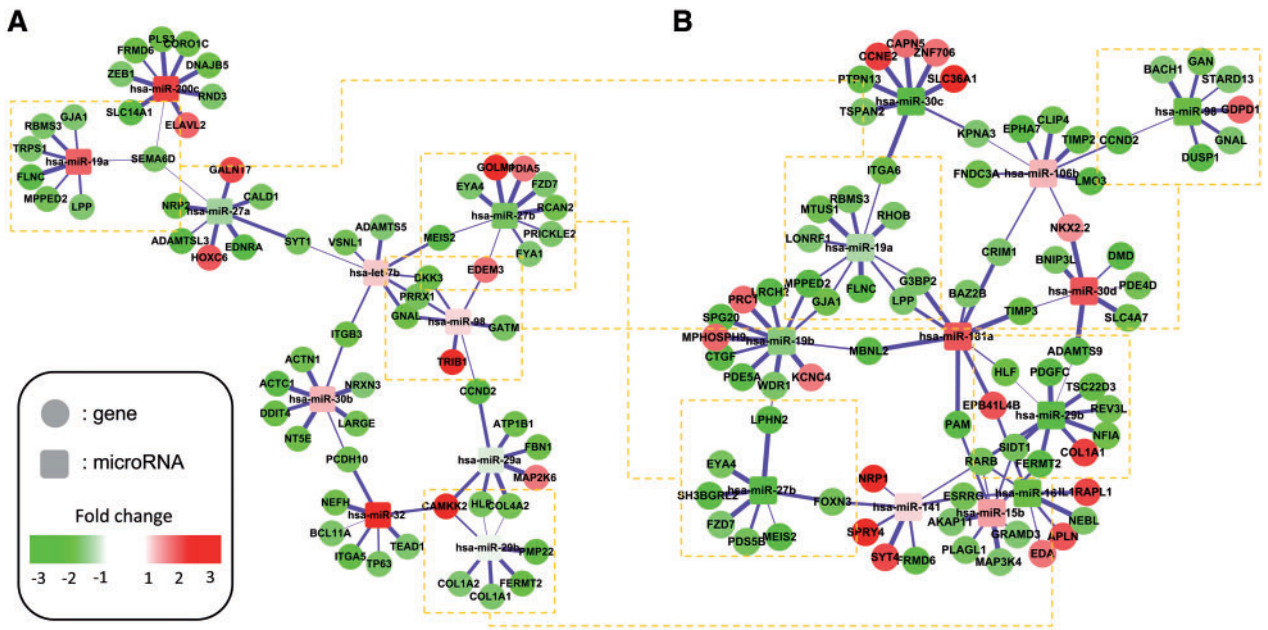

Fig. 3. The PCa miRNA regulatory networks constructed by the Lasso regression model for Normal versus Primaries (**A**) and Primaries versus Metastases (**B**). The four common modules between the two networks are marked by dashed-line boxes. The network visualizations were generated using Cytoscape 2.8 (Smoot *et al.*, 2011).

very robust and informative tool for constructing complex disease-related miRNA regulatory networks, which could provide useful information for the diagnosis and treatment of complex diseases.

We note that the Lasso regression model performed better when using the MSKCC dataset than when using the other two datasets. One of the possible reasons is that we chose the PCa-associated miRNA–mRNA targeting relationships as the known target pair set. The expression levels of both miRNAs and mRNAs in these pairs were significantly altered in the tissue samples, which can greatly reduce the influence of noise in the data. This indicates that the variable dynamic range of miRNA and mRNA expression levels can affect the accuracy of miRNA–mRNA targeting relationship identification, and it suggests that identification methods based on expression data might be further improved if the rates of change in expression levels were considered.

ACKNOWLEDGEMENTS

We thank Dr Xingyi Hang for his helpful discussions and suggestions during the study.

Funding: General Program (30900862, 30900830, 30800196) of the Natural Science Foundation of China; Key and General Program of State Key Laboratory of Proteomics (SKLP-K201004, SKLP-O201002); Special Key Programs for Science and Technology of China (2009ZX09503-002, 2009ZX09301-002, 2009ZX09103-616).

Conflict of Interest: none declared.

REFERENCES

Arvey,A. *et al.* (2010) Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.*, **6**, 363.

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

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Bushati,N. and Cohen,S.M. (2007) MicroRNA functions. *Annu. Rev. Cell Dev. Biol.*, **23**, 175–205.

Doench,J.G. and Sharp,P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.*, **18**, 504–511.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–451.

Eulalio,A. *et al.* (2008) Getting to the root of miRNA-mediated gene silencing. *Cell*, **132**, 9–14.

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

Fu,T.Y. *et al.* (2011) Let-7b-mediated suppression of basigin expression and metastasis in mouse melanoma cells. *Exp. Cell Res.*, **317**, 445–451.

Garzon,R. *et al.* (2009) MicroRNAs in Cancer. *Annu. Rev. Med.*, **60**, 167–179.

Gebeshuber,C.A. *et al.* (2009) miR-29a suppresses tristetraprolin, which is a regulator of epithelial polarity and metastasis. *EMBO Rep.*, **10**, 400–405.

Gennarino,V.A. *et al.* (2009) MicroRNA target prediction by expression analysis of host genes. *Genome Res.*, **19**, 481–490.

Giraldez,A.J. *et al.* (2006) Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, **312**, 75–79.

Gregory,P.A. *et al.* (2008) The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat. Cell Biol.*, **10**, 593–601.

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

Guo,H. *et al.* (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.

Harfe,B.D. (2005) MicroRNAs in vertebrate development. *Curr. Opin. Genet. Dev.*, **15**, 410–415.

Heinzelmann,J. *et al.* (2011) Specific miRNA signatures are associated with metastasis and poor prognosis in clear cell renal cell carcinoma. *World J. Urol.*, **29**, 367–373.

Huang da,W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

John,B. *et al.* (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.

Jones-Rhoades,M.W. *et al.* (2006) MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, **57**, 19–53.

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

Krutzfeldt,J. *et al.* (2005) Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, **438**, 685–689.

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.

Lim,L.P. *et al.* (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.

Liu,J. *et al.* (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, **305**, 1437–1441.

Loayza-Puch,F. *et al.* (2010) Hypoxia and RAS-signaling pathways converge on, and cooperatively downregulate, the RECK tumor-suppressor protein through microRNAs. *Oncogene*, **29**, 2638–2648.

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

Meister,G. *et al.* (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell*, **15**, 185–197.

Nielsen,C.B. *et al.* (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, **13**, 1894–1910.

Olsen,P.H. and Ambros,V. (1999) The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.*, **216**, 671–680.

Rajewsky,N. (2006) MicroRNA target predictions in animals. *Nat. Genet.*, **38** (Suppl. 1), S8–S13.

Ritchie,W. *et al.* (2010) MimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. *Bioinformatics*, **26**, 223–227.

Schmalhofer,O. *et al.* (2009) E-cadherin, beta-catenin, and ZEB1 in malignant progression of cancer. *Cancer Metastasis Rev.*, **28**, 151–166.

Schmitter,D. *et al.* (2006) Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res.*, **34**, 4801–4815.

Sengupta,S. *et al.* (2006) Genome-wide expression profiling reveals EBV-associated inhibition of MHC class I expression in nasopharyngeal carcinoma. *Cancer Res.*, **66**, 7999–8006.

Sethupathy,P. *et al.* (2006a) TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.

Sethupathy,P. *et al.* (2006b) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods*, **3**, 881–886.

Sengupta,S. *et al.* (2008) MicroRNA 29c is down-regulated in nasopharyngeal carcinomas, up-regulating mRNAs encoding extracellular matrix proteins. *Proc. Natl Acad. Sci. USA*, **105**, 5874–5878.

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

Slaby,O. *et al.* (2010) Expression of miRNA-106b in conventional renal cell carcinoma is a potential marker for prediction of early metastasis after nephrectomy. *J. Exp. Clin. Cancer Res.*, **29**, 90.

Smoot,M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Stanhope,S.A. *et al.* (2009) Statistical use of argonaute expression and RISC assembly in microRNA target identification. *PLoS Comput. Biol.*, **5**, e1000516.

Steele,R. *et al.* (2010) MBP-1 upregulates miR-29b that represses Mcl-1, collagens, and matrix-metalloproteinase-2 in prostate cancer cells. *Genes Cancer*, **1**, 381–387.

Sun,Q. *et al.* (2010) Hsa-mir-27a genetic variant contributes to gastric cancer susceptibility through affecting miR-27a and target gene expression. *Cancer Sci.*, **101**, 2241–2247.

Swami,M. (2010) Small RNAs: targeting transcripts for destruction. *Nat. Rev. Genet.*, **11**, 672.

Takeshita,F. *et al.* (2010) Systemic delivery of synthetic microRNA-16 inhibits the growth of metastatic prostate tumors via downregulation of multiple cell-cycle genes. *Mol. Ther.*, **18**, 181–187.

Taylor,B.S. *et al.* (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell*, **18**, 11–22.

Team,R. (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Vallejo,D.M. *et al.* (2011) Targeting Notch signalling by the conserved miR-8/200 microRNA family in development and cancer cells. *EMBO J.*, **30**, 756–769.

Volinia,S. *et al.* (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl Acad. Sci. USA*, **103**, 2257–2261.

Volinia,S. *et al.* (2010) Identification of microRNA activity by Targets' Reverse EXpression. *Bioinformatics*, **26**, 91–97.

Wang,Y. *et al.* (2011) Transforming growth factor-beta regulates the sphere-initiating stem cell-like feature in breast cancer through miRNA-181 and ATM. *Oncogene*. **30**, 1470–80

Wightman,B. *et al.* (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, **75**, 855–862.

Yao,J. *et al.* (2010) MicroRNA-30d promotes tumor invasion and metastasis by targeting Galphai2 in hepatocellular carcinoma. *Hepatology*, **51**, 846–856.