Journal of Theoretical Biology 337 (2013) 54–60

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

RRSM with a data-dependent threshold for miRNA target prediction

Wan J. Hsieh, Hsiuying Wang *

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

HIGHLIGHTS

- Predicting miRNA target genes is one of the important issues in bioinformatics.
- The RRSM has been proposed for miRNA target prediction in the literature.
- RRSM with a data-dependent threshold is proposed in this study.
- The new method can select more experimentally validated targets than RRSM.

ARTICLE INFO

Article history:
Received 22 May 2013
Received in revised form
26 July 2013
Accepted 1 August 2013
Available online 13 August 2013

Keywords:
The relative *R* squared method
Correlation analysis
Regression model
p-value

ABSTRACT

Predicting miRNA target genes is one of the important issues in bioinformatics. The correlation analysis is a widely used method for exploring miRNA targets through microarray data. However, the experimental results show that correlation analysis leads to large false positive or negative results. In addition, the correlation analysis is not appropriate when multiple miRNAs simultaneously regulate a gene. Recently, the relative *R* squared method (RRSM) has been proposed for miRNA target prediction, which is shown to be superior to some existing methods. To adopt the RRSM, we need first to set thresholds to select a proportion of potential targets. In the previous studies, the threshold is set to be fixed, which does not depend on the characteristic of a gene. Due to the diversity of the functions of genes, a data-dependent threshold may be more feasible in real data applications than a data-independent threshold. In this study, we propose a threshold selection method which is based on the distribution of the relative *R* squared statistic. The proposed method is shown to significantly improve the previous prediction results by selecting more experimentally validated targets.

1. Introduction

Predicting target genes is one of the important research topics in bioinformatics, such as discovering microRNA (miRNA) interactions or transcription factor binding sites. Recent works have revealed that miRNAs play important roles in various biological processes (Bartel, 2004; Ambros, 2004; Broderick and Zamore, 2011). In the previous study, the correlation analysis is a widely used method for exploring target genes of a miRNA through microarray data (van Dongen et al., 2008; Bartonicek and Enright, 2010). However, experimental results show that correlation analysis does not lead to accurate results (Huang et al., 2007a, 2007b; Wang and Li, 2009; Hsieh and Wang, 2011). These previous studies indicated that for many miRNAs, the correlation coefficient of the microarray expression of a miRNA and that of its confirmed target is nearly zero. When the correlation coefficient is not high, it is hard to use any standard statistical approaches to explore miRNA targets because there are no significant statistical evidence for a relationship between a miRNA and its true targets in terms of the conventional statistical methods. In addition, the correlation analysis is not appropriate to be used when multiple factors simultaneously function on a target. In many biological applications, it is more appropriate to build a statistical model, such as a regression model, than using the correlation analysis to analyze the data (Wang and Li, 2009; Lu and Wang, 2012).

Recently, the relative *R* squared method (RRSM), which is developed based on a regression model, has been proposed for target gene prediction, and it is shown to be superior to some existing methods (Wang and Li, 2009; Hsieh and Wang, 2011; Wang et al., 2011). RRSM is proposed to analyze data from a relative instead of from the absolute statistical point of view. In biological systems, it is usual that a gene is simultaneously regulated by multiple miRNAs. To describe the relationship between the expression profiles of miRNAs and their target genes, we are interested in exploring a statistical model to capture the relationship. With this estimated statistical model, we can predict potential target genes of a miRNA for further experimental validation. Due to the high cost of experimentation, we expect to find a reasonable amount of potential targets for further experimental validation in finding the true targets. Therefore, establishing

* Corresponding author. Tel.: +886 3 571 2121x56813; fax: +886 3 572 8745.
E-mail address: wang@stat.nctu.edu.tw (H. Wang).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.08.002

an efficient and simple method to reduce the false discovery rate or negative rate of the target prediction is an essential issue. In addition to predicting miRNA targets, many studies focus on constructing miRNA-regulated gene networks to explore miRNA-mRNA regulatory relatinships such as CoMeTa tool (Gennarino et al., 2012; Le et al., 2013). In this study, we do not deeply discuss the network analysis because we manily focus on the target precition problem.

Since the true biological model, which can capture the expression data relationship between target genes and miRNAs, may be very complicated, it is hard to build the true model. A feasible way is to approximate the relationship by a linear regression model although a linear model may not really well fit the data. In a regression model, the coefficient of determination, denoted as R^2, with value between 0 and 1 is a criterion used to evaluate the fitness of the model to the data (Buse, 1973; Cameron and Windmeijer, 1997). A model with a larger R^2 is preferable to be used to fit the data. Since in real applications, the biological relationship cannot be characterized by a linear function, the R^2 based on a linear model to fit the data might be low. RRSM, which is proposed to overcome this disadvantage of the R^2 criterion, is successfully used to predict potential targets. Nevertheless, the threshold selection is a main issue in adopting RRSM to select the potential targets. The false discovery rate and the false negative rate of the prediction results strongly depend on the threshold selection. To provide a more depth investigation of the threshold selection, in this study, we focus on exploring the theoretical property of the RRSM, and then we base on the established property to propose a more reliable method to select the thresholds of RRSM.

In the previous studies, the fixed threshold criterion was adopted in RRSM (Wang and Li, 2009; Hsieh and Wang, 2011; Wang et al., 2011). The procedure of RRSM is to compare two different R^2 values with respect to two different linear models. We call that one is a full model and the other one is a reduced model. The explanatory variables in the reduced model are in a subset of the explanatory variables in the full model. The ratio of the R^2 value with respect to the reduced model to the R^2 value of the full model is a relative R squared value. When the relative R squared value is greater than a threshold, we select the targets corresponding to the reduced model as the potential targets. For a miRNA or a transcription factor, to predict target genes, Wang and Li (2009), Hsieh and Wang (2011) and Wang et al., (2011) used the same threshold for the relative R squared value when building regression models for different genes. It is worth noting that in these studies although the goal is to find the target genes of a miRNA, the RRSM is to build a regression model for each gene with gene expression values as the response variables and the miRNA expression values as explanatory variables. The reason is that the expression of a gene may be regulated by particular miRNAs, but it is not that the expression of a miRNA is regulated by particular genes. Therefore, a regression model is build for each gene with different miRNAs as explanatory variables. In the previous studies, the threshold is set to be the same (fixed) for each regression model, which does not depend on the characteristic of a gene (Wang and Li, 2009; Hsieh and Wang, 2011). In this study, we propose a data-dependent threshold selection method based on the distribution of the relative R squared statistic, which is shown to significantly improve the prediction results of RRSM with a fixed (data-independent) threshold criterion from a simulation study and miRNA data analysis.

2. Results

In this section, we review the RRSM procedure with a data-independent threshold, and propose the procedure for RRSM with a data-dependent threshold.

2.1. Matrix form for RRSM

The datasets we used in this study are the mRNA and miRNA expression data for 114 human miRNAs and 16 063 mRNAs across a mixture of 88 normal and cancerous tissue samples common to the two datasets used in Huang et al. (2007a) and Hsieh and Wang (2011). To investigate the theoretical property of the relative R squared method, we represent the relative R squared method in Wang and Li (2009) with a matrix form.

Let y_j denote the expression data of a mRNA in the jth tissue and let x_{ji} denote the expression data of the ith miRNA in the jth tissue, where $j=1,\ldots,n$ and $i=1,\ldots,p$.

Full model (Ω):

$$y_j = b_0x_{j0} + b_1x_{j1} + b_2x_{j2} + \cdots + b_px_{jp} + \varepsilon_j, \quad j = 1,2,\ldots,n$$

or

$$\mathbf{Y} = \mathbf{X}_\Omega\beta_\Omega + \varepsilon \tag{1}$$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^T$ is the response variable and $\mathbf{x_i} = (x_{1i}, \ldots, x_{ni})^T$, $i = 1,\ldots,p$ is the ith explanatory variable and $\mathbf{x_0} = (x_{10}, \ldots, x_{n0})^T = (1,\ldots,1)^T$ is a constant term. $\beta_\Omega = (b_0, \ldots, b_p)$ are regression parameters, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the error term distributed as a multivariate normal distribution $N(0,\sigma^2 I_n)$. Let $\mathbf{x_i} = (x_{1i}, \ldots, x_{ni})^T$, $i = 1,\ldots,p$ be the ith explanatory variable and $\mathbf{x_0} = (x_{10}, \ldots, x_{n0})^T = (1,\ldots,1)^T$ be a constant term. Under the model (1), the least squared estimator for β_Ω is $\hat{\beta}_\Omega = (\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_p)^T = (\mathbf{X}_\Omega^T\mathbf{X}_\Omega)^{-1}\mathbf{X}_\Omega^T\mathbf{Y}$, and let $\hat{\mathbf{Y}}_\Omega = \mathbf{X}_\Omega\hat{\beta}_\Omega$. The R^2 value of the model (1) is defined as $R^2_\Omega = SSR_\Omega/SST$, where $SST = ||\mathbf{Y}-\overline{Y}||^2$ is the total sum of squares, $SSR_\Omega = ||\hat{\mathbf{Y}}_\Omega-\overline{Y}||^2$ is the regression sum of squares and $\overline{Y}$ is the mean of $y_1, y_2, \ldots, y_n$. The goal of RRSM is to find high-confidence explanatory variables such that it can significantly affect the response variables. The first step of RRSM is to find p-values for testing $H_{0_i} : b_i = 0$, $i = 1,\ldots,p$. For a fixed i, the p-value for testing the null hypothesis based on the estimator $\hat{\beta}_\Omega$ is defined

$$Pr(|W| \geq \hat{b}_i/\sqrt{var(\hat{b}_i)}), \tag{2}$$

where W denotes the t distribution with degrees of freedom $n-p-1$ and $var(\hat{b}_i)$ denotes the variance of the estimator $\hat{b}_i$ (Wang and Li, 2009). Note that $var(\hat{b}_i)$ can be approximated by the ith diagonal element of $(\mathbf{X}_\Omega{}^T\mathbf{X}_\Omega)^{-1}\hat{\sigma}^2$ which is due to the fact that the estimator $\hat{\beta}_\Omega$ is distributed as a normal distribution $N(\beta_\Omega, (\mathbf{X}_\Omega{}^T\mathbf{X}_\Omega)^{-1}\sigma^2)$ and $\hat{\sigma}^2 = ||\mathbf{Y}-\hat{\mathbf{Y}}_\Omega||^2/(n-p-1)$ is an estimator of σ^2. Here, we set a threshold p_0 and select an explanatory variable $\mathbf{x_i}$ as a potential explanatory variable if the corresponding p-value for testing the null hypothesis $H_{0_i} : b_i = 0$ is less than threshold p_0. Assume that there are k ($k \leq p$) variables $\{\mathbf{x}_{\eta_1}, \mathbf{x}_{\eta_2}, \ldots, \mathbf{x}_{\eta_k}\}$, $\eta_1 < \eta_2 < \cdots < \eta_k$ which have been selected by the p-value criterion. Then we rebuild the regression model using these k explanatory variables as follows.

Reduced model (ω):

$$y_j = b^*_0x_{j\eta_0} + b^*_1x_{j\eta_1} + b^*_2x_{j\eta_2} + \cdots + b^*_kx_{j\eta_k} + \varepsilon^*_j, \quad j = 1,2,\ldots,n$$

or

$$\mathbf{Y} = \mathbf{X}_\omega\beta_\omega + \varepsilon^* \tag{3}$$

where $\mathbf{x}_{\eta_0} = (x_{1\eta_0}, \ldots, x_{n\eta_0})^T = (1,\ldots,1)^T$ is the constant term and $\varepsilon^* = (\varepsilon^*_1, \ldots, \varepsilon^*_n)^T$ is the error term distributed as a multivariate normal distribution $N(0,\sigma^2 I_n)$. In model (3), the least squared error estimator is $\hat{\beta}_\omega = (\hat{b}^*_0, \hat{b}^*_1, \ldots, \hat{b}^*_k)^T = (\mathbf{X}^T_\omega\mathbf{X}_\omega)^{-1}\mathbf{X}^T_\omega\mathbf{Y}$, where $\mathbf{X}_\omega = (x_{j\eta_i})_{n\times(k+1)}$. Let $\hat{\mathbf{Y}}_\omega = \mathbf{X}_\omega\hat{\beta}_\omega$. We calculate the R^2 value with respect to model (3), say R^2_ω, where $R^2_\omega = SSR_\omega/SST$ and $SSR_\omega = ||\hat{\mathbf{Y}}_\omega-\overline{Y}||^2$. The ratio of R^2_ω to R^2_Ω, R^2_ω/R^2_Ω, which is defined as the relative R squared value (Wang and Li, 2009). Then we set a threshold for R^2_ω/R^2_Ω, say s. If R^2_ω/R^2_Ω is larger than s, the variables $\{\mathbf{x}_{\eta_1}, \mathbf{x}_{\eta_2}, \ldots, \mathbf{x}_{\eta_k}\}$ are selected. Otherwise, we do not select any variable. Since RRSM considers the criterion of the ratio of two

R^2 values, it adopts a relative statistical viewpoint instead of an absolute statistical viewpoint to select high-confidence explanatory variables. It is worth mentioning that RRSM is not only applied to the linear regression but it also can be applied to more complicated models.

2.2. Data-dependent threshold selection

In the previous microarray data studies for RRSM, for each gene, the threshold (p_0, s) in selecting targets of the regression model is set to be fixed. That is, p_0 and s are set to be the same for all genes. However, a more flexible criterion is to set the threshold such that it depends on the data, i.e. the expression profile of the gene. But how to set a reasonable data-dependent threshold is a challenging task. In this study, we investigate the distribution of the relative R squared statistic and use it to derive a data-dependent threshold for RRSM. The details for the theoretical derivation of the RRSM procedure with a data-dependent threshold are given in the Method section. The steps of a data-dependent threshold for RRSM are briefly described in Procedure 1 and Fig. 1.

Procedure 1:. The steps of RRSM with a data-dependent threshold.

Step 1: Build a regression model (1).
Step 2: Set p_0 as a critical point of p-value. The variable $\mathbf{x_i}$ is selected if $Pr(|W| \geq \hat{b}_i / \sqrt{var(\hat{b}_i)}) \leq p_0$, and the selected variables are denoted as $\{\mathbf{x}_{\eta_1}, \mathbf{x}_{\eta_2}, \ldots, \mathbf{x}_{\eta_k}\}$.

Step 3: Base on $\{\mathbf{x}_{\eta_1}, \mathbf{x}_{\eta_2}, \ldots, \mathbf{x}_{\eta_k}\}$ to build a new regression model (3).

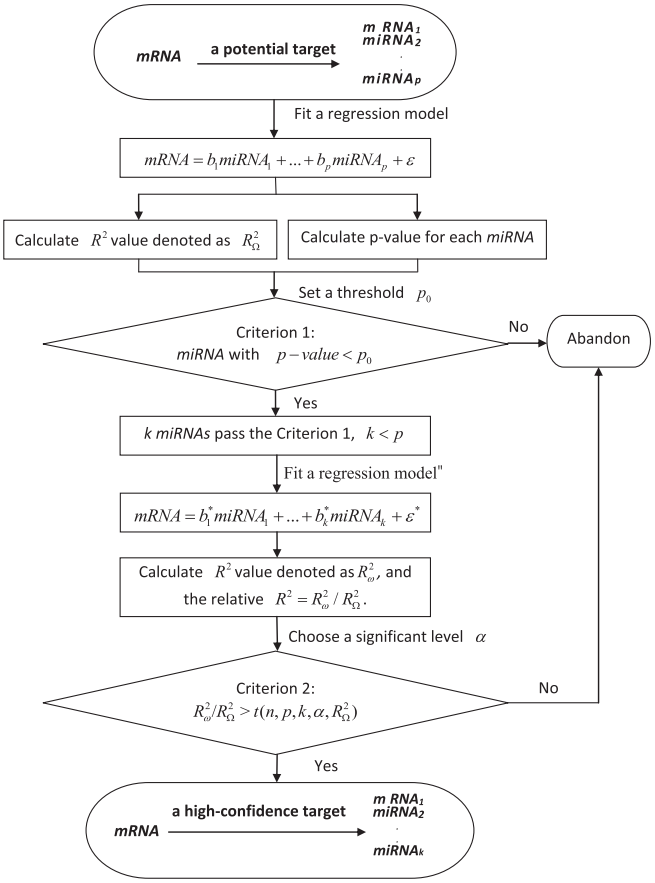

Fig. 1. The flowchart of the procedure for a data-dependent threshold for RRSM.

Step 4: Calculate the R squared values of model (1) and (3), say R^2_{Ω} and R^2_{ω}, respectively, and calculate the relative R squared value, $R^2_{\omega}/R^2_{\Omega}$.
Step 5: Choose a significant level α. The variables $\{\mathbf{x}_{\eta_1}, \mathbf{x}_{\eta_2}, \ldots, \mathbf{x}_{\eta_k}\}$ are regarded as significant variables to affect the response variable if $R^2_{\omega}/R^2_{\Omega} > t(n, p, k, \alpha, R^2_{\Omega})$,
where

$$t(n, p, k, \alpha, R^2_{\Omega}) = 1 - (F^{\alpha}_{p-k, n-p-1} \times (p-k))/(n-p-1) \times (1/R^2_{\Omega} - 1). \quad (5)$$

and $F^{\alpha}_{\nu_1, \nu_2}$ denotes the upper α cut of point of a F distribution with degrees of freedom ν_1 and ν_2.

The difference between the RRSM with a fixed threshold and RRSM with a data-dependent threshold is that the former is to set a constant threshold s and the latter adopts a data-dependent threshold $t(n, p, k, \alpha, R^2_{\Omega})$ which depends on the R^2_{Ω} value. The code for RRSM with a data-dependent threshold can be download at http://www.stat.nctu.edu.tw/~hwang/website_wang%20new.htm. In addition, a total of 1536 high-confidence targets (Table S1 in the supplementary materials) were discovered in this study and we list targets associating with corresponding p-values. It is worth noting that there are 205 high-confidence targets with p-value less than 0.05 which can be ranked to be more potential targets than the other 1331 selected targets.

3. Simulation

To evaluate the performance of RRSM with a data-dependent threshold, we conduct a simulation study to compare it with the RRSM proposed in the previous study. The two methods are evaluated in terms of their false negative rates and false positive rates. The false negative rate of a method is defined as v_1/h, where h denotes the number of true explanatory variables, and v_1 is the number of the true explanatory variables which are not selected by the method among the h true explanatory variables. The false positive rate of a method is defined to be the ratio of $v_2/(r-h)$, where r denotes the number of all explanatory variables, and v_2 denotes the number of variables selected by this method which are not the true explanatory variables.

The steps of the simulation study are described as follows. We first generate h explanatory variables $\mathbf{x_i} = (x_{1i}, \ldots, x_{ni})^T$, $i = 1, 2, \ldots, h$, and then generate a sample y_j, $j = 1, \ldots, n$ from a given linear model (3) based on these h explanatory variables. After that, we generate $r-h$ noise explanatory variables, $\mathbf{x^*_i} = (x^*_{1i}, \ldots, x^*_{ni})^T$, $i = 1, \ldots, r-h$. With the sample $y_1, \ldots, y_n$, $\{\mathbf{x_1}, \ldots, \mathbf{x_h}\}$ and $\{\mathbf{x^*_1}, \ldots, \mathbf{x^*_{r-h}}\}$, we follow the RRSM to select explanatory variables. After that we calculate the false negative rates and the false positive rates of the two methods. We replicate 100 simulation process, and then calculate the average of their false negative rates and the false positive rates.

We present the simulation results for the cases of varying a term of n, p or k when the other two terms are fixed. Figs. 2–4 present the simulation results of the false negative rates and the false positive rates of the two methods for the cases of different n, p or k, respectively. In this study, the threshold s for RRSM with a fixed threshold is set to be 0.95 and the significant level α for RRSM with a data-dependent threshold is set to be 0.05. The threshold p_0 for both RRSMs are selected to be around 1/3 for the most cases or in the range of (0.1,0.6).

Fig. 2 shows that results of $n = 50, 100, 150, 200, 250$ and 300 when $p = 15$ and $k = 5$. The false positive rates and the false negative rates of RRSM with a fixed threshold are significantly larger than those of RRSM with a data-dependent threshold. In addition, the false positive rates and false negative rates both

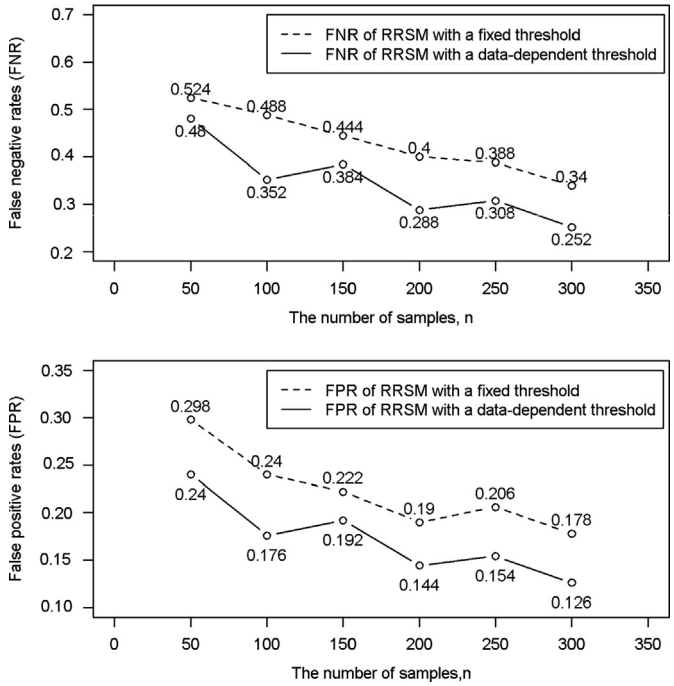

Fig. 2. The false negative rates (dashed lines) and false positive rate (solid lines) of two RRSMs for $p=15$, $k=5$ and $n=50$, 100, 150, 200, 250 and 300.

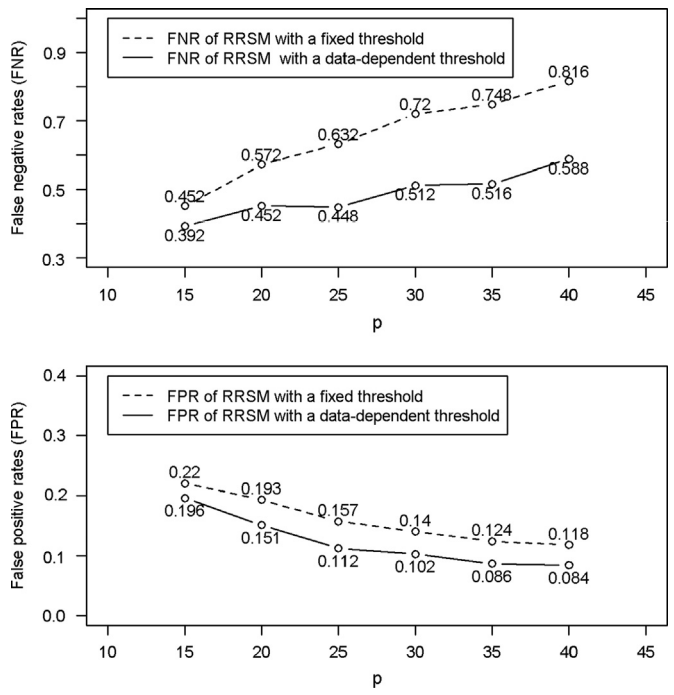

Fig. 3. The false negative rates (dashed lines) and false positive rate (solid lines) of two RRSMs for $n=100$, $k=5$ and $p=15$, 20, 25, 30, 35 and 40.

decrease as the sample size increases. It reveals that the false positive rates and the false negative rates of RRSM can be reduced when the sample size is large enough.

Fig. 3 presents the results for different p when n and k are 100 and 15, respectively. The comparison results are the same as the case in Fig. 2 that both false rates of RRSM with a data-dependent threshold are lower than those of RRSM with a fixed threshold. It is worth noting that the false negative rate increases as the number of variables, r, increases. It is due to the fact that noise level increases when r increases, but h is fixed. Since the noise level increases, the chance of the true explanatory variables which are not selected by the method increases. This fact might cause the increase of the false negative rate.

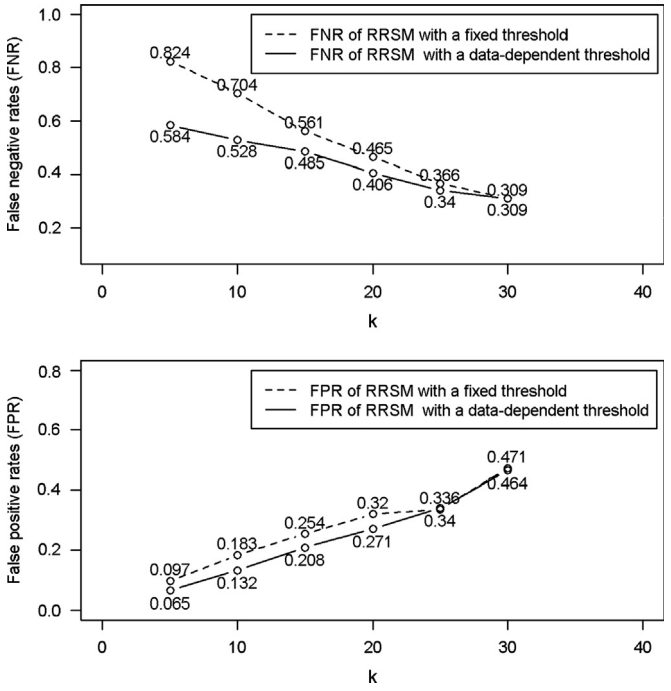

Fig. 4. The false negative rates (dashed lines) and false positive rate (solid lines) of two RRSMs for $n=100$, $p=50$ and $k=5$, 10, 15, 20, 25 and 30.

The results for the last case for different k are presented in Fig. 4. The RRSM with a data-dependent threshold still outperforms the RRSM with a fixed threshold. It is worth noting that the false negative rate decreases as the number of true explanatory variables h increases. It is due to the fact that the noise level decreases when h increases, but r is fixed. Since the noise level decreases, the number of the selected variables which are the not true explanatory variables might decrease. In this case, v_1 decreases and h increases which causes the decrease of the false negative rate. The above simulation results reveal that RRSM with a data-dependent threshold is a more feasible approach for finding true explanatory variables.

4. Data analysis

In this section, we analyze microarray expression profiling of miRNA and mRNA in Huang et al. (2007a) and Hsieh and Wang (2011) to show the advantage of RRSM with a data-dependent threshold. The goal of analyzing these datasets is to find the mRNAs regulated by a specific miRNA. Since there are too many possible pairs ($114\times$ 16 063 pairs) of miRNA and mRNA among 114 human miRNAs and 16 063 mRNAs across 88 tissues, it is inefficient to directly consider all possible pairs. There are 6387 potential target pairs filtered by TargetScanS (Huang et al., 2007a; Hsieh and Wang, 2011). TargetScanS is a prediction tool by searching for conserved 8mer and 7mer sites that match the seed region of miRNAs (Grimson et al., 2007; Lewis et al., 2005; Friedman et al., 2009). Although the potential target pairs are reduced to 6387 pairs, there are still many false positive targets in the 6387 pairs. Hsieh and Wang (2011) adopted RRSM with a fixed threshold to find the high-confidence interactions. In this study, we apply the RRSM with a data-dependent threshold to find the high-confidence interactions and compared the result with Hsieh and Wang (2011).

To compare the two methods, by the criterion used in Hsieh and Wang (2011), we investigate the accuracies of both methods in terms of the number of experimentally validated genes being selected. The experimentally validated genes can be obtained from TarBase (Papadopoulos et al., 2009) and miRTarBase (Hsu et al., 2011). To compare the performances of these methods, we examine the accuracy of these methods by comparing their selected target genes with the experimentally confirmed genes in TarBase and miRTarBase. For the 6387 potential targets, there are only 24 and 138 confirmed genes in TarBase and miRTarBase, respectively.

To compare the methods, we use these criteria to select one-fourth of genes (around 1600 genes) from the 6387 potential targets because Huang et al. (2007a,b), Hsieh and Wang (2011) and Wang and Li (2009) compared different methods using one-fourth of genes. To make a more objective comparison, the results for different thresholds of each method such that each of them has better performance are presented.

Table 1 presents the results for different methods under different thresholds. There are 8–10 interactions (experimentally validated genes) in TarBase selected by RRSM with a fixed threshold, and there are 11–13 interactions in TarBase selected by RRSM with a data-dependent threshold. RRSM with a data-dependent threshold selects more interactions than RRSM with a fixed threshold. In addition, we also compare the selecting results through the database miRTarBase. There are 20–26 interactions selected by RRSM with a fixed threshold, and 28–29 interactions selected by RRSM with a data-dependent threshold, respectively. It shows that RRSM with a data-dependent threshold is better than RRSM with a fixed threshold applying in these datasets for predicting high-confidence targets.

Furthermore, to make more comprehensive comparisons for the two RRSMs, we apply these two methods to another miRNA target prediction tool by Huang et al. (2007a) and mouse expression profiles. Huang et al. (2007a) adopted a prediction tool, GenMiR++, to select the 1597 high-confidence targets among the same data set used in the above analysis. There are only four interactions in TarBase among these 1597 high-confidence targets. We apply the RRSM with a fixed threshold and RRSM with a data-dependent threshold to 1597 high-confidence targets, respectively. Table 2 shows the comparison results for different thresholds. To cover the four interactions in TarBase, we found that the RRSM with a fixed threshold has to select more high-confidence targets than those selected by the RRSM with a data-dependent threshold. Table 2 shows that the selected numbers of high-confidence targets for RRSM with a data-dependent threshold are 617, 662 and 680 for some thresholds. However, the RRSM with a fixed threshold is necessary to select 802, 810 and 883 high-confidence targets to cover these four interactions. Thus, the RRSM with a data-dependent threshold has lower false discovery rate than the RRSM with a data-independent threshold for this data anaysis.

In addition to analyzing human miRNA expression profiles in the above, we apply the RRSMs to mouse data. Wang and Li (2009) proposed the RRSM with a fixed threshold for mouse miRNA target prediction and found that there are two Tarbase interactions in 1770 potential targets; the relationship between miR-181a and BCL2 mRNA and the relationship between miR-181a and HOXA11 mRNA. They used thresholds $p_0 = 0.47$ and $s = 0.995$, which leads to 448 high-confidence targets with one interaction in TarBase. For covering these two interactions in TarBase, they relaxed thresholds to $p_0 = 0.67$ and $s = 0.9999$, which leads to 715 high-confidence targets. However, we adopt the RRSM with a data-dependent threshold by using different thresholds which lead to 177, 236 and 260 high-confidence targets with these two interactions in TarBase. The comparisons between the two methods for mouse data set are presented in Table 3.

In summary, we apply the RRSM with a fixed threshold and RRSM with a data-dependent threshold in different datasets and prediciton methods. These results reveal that RRSM with a data-dependent threshold is more powerful than the RRSM with a fixed threshold for predicting high-confidence targets.

5. Discussion

In this study, we propose a data-dependent threshold selection method based on the distribution of the relative R squared statistic to provide a feasible rule for selecting useful thresholds for the RRSM method. The thresholds of RRSM are set to be the same for each data in the previous study, which do not depend on the characteristic of a gene. We show that a data-dependent threshold criterion leads to more convincing results in the simulation study and real data analysis. From the data analysis, the RRSM with the data-dependent threshold is shown to predict more valid targets than RRSM with a fixed threshold. Therefore, we conclude that the proposed threshold selection criterion can benefit the variable selection in biology or other related analysis.

6. Methods

In order to derive our main theorem, we recall some useful properties of a linear model and define some notations first. We consider a simpler case to simplify the model assumptions that let $\{\mathbf{x}_{\eta_0}, \mathbf{x}_{\eta_1}, \ldots, \mathbf{x}_{\eta_k}\}$ in reduced model (3) be the first $k+1$ columns of $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p\}$ in full model (1). We consider a null hypothesis

Table 1
Interaction numbers in TarBase and miRTarBase selected by RRSM with a fixed threshold or RRSM with a data-dependent threshold for human expression profiles.

Method	Number of high-confidence targets	Number of interactions in TarBase	Number of interactions in miRTarBase
RRSM with a fixed threshold			
$s = 0.995$, $p_0 = 0.77$	1559	10	26
$s = 0.995$, $p_0 = 0.75$	1342	9	26
$s = 0.990$, $p_0 = 0.72$	1485	8	25
$s = 0.950$, $p_0 = 0.60$	1388	8	20
$s = 0.900$, $p_0 = 0.57$	1519	8	20
RRSM with a data-dependent threshold			
$\alpha = 0.99$, $p_0 = 0.583$	1536	13	29
$\alpha = 0.88$, $p_0 = 0.4$	1568	11	28

Table 2
The high-confidence target numbers selected by two RRSMs in applying to the predicted results by GenMiR++ for covering four interactions in TarBase under different thresholds.

Method	Number of high-confidence targets	Number of interactions in TarBase
GenMiR++	1597	4
RRSM with a fixed threshold		
$s = 0.99$, $p_0 = 0.72$	802	4
$s = 0.995$, $p_0 = 0.77$	810	4
$s = 0.99$, $p_0 = 0.75$	883	4
RRSM with a data-dependent threshold		
$\alpha = 0.97$, $p_0 = 0.72$	617	4
$\alpha = 0.99$, $p_0 = 0.75$	662	4
$\alpha = 0.995$, $p_0 = 0.77$	680	4

Table 3
High-confidence target numbers selected by RRSM with a fixed threshold and RRSM with a data-dependent threshold for covering interactions in TarBase for mouse expression profiles.

Method	Number of high-confidence targets	Number of interactions in TarBase
RRSM with a fixed threshold		
$s=0.995$, $p_0=0.47$	448	1
$s=0.9999$, $p_0=0.67$	715	2
RRSM with a data-dependent threshold		
$\alpha=0.5$, $p_0=0.1$	177	2
$\alpha=0.3$, $p_0=0.1$	236	2
$\alpha=0.5$, $p_0=0.15$	260	2

$H_0: b_{k+1}=b_{k+2}=\cdots=b_p=0$. Then under the null hypothesis, the full model (1) is the reduced model (3). The hypothesis can be expressed as $H_0: \mathbf{A}\beta_\Omega=0$, where $A=(a_{ij})_{(p-k)\times(p+1)}$ is a $(p-k)\times(p+1)$ matrix, and

$$a_{ij}=\begin{cases}1 & j=i+k+1 \text{ for } i=1,2,\ldots,p-k\\ 0 & \text{otherwise}\end{cases}$$

Before giving our main theorem, we introduce a useful corollary (Seber and Lee, 2003, Theorem 4.1).

Corollary 1. *To test $H_0: \mathbf{A}\beta_\Omega=0$, under H_0, the statistic*

$$((SSE_\omega-SSE_\Omega)\times(n-p-1))/(SSE_\Omega\times(p-k))$$

is distributed as $F_{p-k,n-p-1}$, where $F_{p-k,n-p-1}$ denotes the F distribution with $p-k$ and $n-p-1$ degrees of freedom, and $SSE_\Omega=\|\mathbf{Y}-\hat{\mathbf{Y}}_\Omega\|^2$ and $SSE_\omega=\|\mathbf{Y}-\hat{\mathbf{Y}}_\omega\|^2$ are the residual sums of squares of the full model (1) and reduced model (3), respectively.

Based on Corollary 1, we derive the distribution of the relative R squared statistic in Theorem 1.

Theorem 1. *Under models (1) and (3), for $k<p$, to test*

$$H_0: \mathbf{A}\beta_\Omega=0 \quad vs. \quad H_1: \mathbf{A}\beta_\Omega\neq 0,$$

we have

$$((R^2_\Omega-R^2_\omega)\times(n-p-1))/((1-R^2_\Omega)\times(p-k))\sim F_{p-k,n-p-1}.$$

Proof. By definition and properties of linear regression model (Seber and Lee, 2003), we have

$$R^2_\Omega=SSR_\Omega/SST \quad \text{and} \quad R^2_\omega=SSR_\omega/SST$$

and

$$SSR_\Omega=SST-SSE_\Omega \quad \text{and} \quad SSR_\omega=SST-SSE_\omega.$$

The numerator of $(R^2_\Omega-R^2_\omega)/(1-R^2_\Omega)$ can be rewritten as

$$R^2_\Omega-R^2_\omega=(SSR_\Omega-SSR_\omega)/SST=((SST-SSE_\Omega)-(SST-SSE_\omega))/SST$$
$$=(SSE_\omega-SSE_\Omega)/SST$$

and the denominator of $(R^2_\Omega-R^2_\omega)/(1-R^2_\Omega)$ can be rewritten as

$$1-R^2_\Omega=1-SSR_\Omega/SST=(SST-SSR_\Omega)/SST=SSE_\Omega/SST.$$

Consequently, we have

$$(R^2_\Omega-R^2_\omega)/(1-R^2_\Omega)=((SSE_\omega-SSE_\Omega)/SST)/(SSE_\Omega/SST)=(SSE_\omega-SSE_\Omega)/SSE_\Omega.$$

If both the numerator and the denominator of $(SSE_\omega-SSE_\Omega)/SSE_\Omega$ are divided by σ^2, we have

$$(SSE_\omega-SSE_\Omega)/\sigma^2\sim\chi^2_{[n-(k+1)]-[n-(p+1)]}\equiv\chi^2_{p-k}$$

and

$$SSE_\Omega/\sigma^2\sim\chi^2_{n-p-1},$$

where χ^2_ξ denotes the chi squared distribution with ξ degrees of freedom (Seber and Lee, 2003). Under H_0, by Corollary 1 we have

$$((SSE_\omega-SSE_\Omega)/\sigma^2)/(SSE_\Omega/\sigma^2)\times(n-p-1)/(p-k)$$
$$=(SSE_\omega-SSE_\Omega)/SSE_\Omega\times(n-p-1)/(p-k)\sim F_{p-k,n-p-1}.$$

Then the proof is complete.

We will use the property in Theorem 1 to derive a data-dependent threshold for RRSM. Note that $(R^2_\Omega-R^2_\omega)/(1-R^2_\Omega)$ can be written as

$$(R^2_\Omega-R^2_\omega)/(1-R^2_\Omega)=(1-R^2_\omega/R^2_\Omega)/(1/R^2_\Omega-1),$$

where R^2_ω/R^2_Ω is the relative R squared value. Define

$$F=(1-R^2_\omega/R^2_\Omega)/(1/R^2_\Omega-1)\times(n-p-1)/(p-k)$$

According to Theorem 1, to test $H_0: \mathbf{A}\beta_\Omega=0$, we calculate F value and reject H_0 at level α of significance if

$$F\geq F^\alpha_{p-k,n-p-1}, \tag{4}$$

For a fixed R^2_Ω value, a larger relative R squared value leads to a smaller F value, which results that we do not reject the hypothesis when the relative R squared value is large. Therefore, by rewritting Eq. (4), we obtain a test based on the relative R squared statistic in Theorem 2.

Theorem 2. *Under models (1) and (3), for $k<p$, to test*

$$H_0: \mathbf{A}\beta_\Omega=0 \quad vs. \quad H_1: \mathbf{A}\beta_\Omega\neq 0,$$

a level α test based on the critical (reject) region derived by (4) is

$$\varphi=\begin{cases}1 & \text{if } R^2_\omega/R^2_\Omega\leq t(n,p,k,\alpha,R^2_\Omega)\\ 0 & \text{otherwise}\end{cases},$$

where

$$t(n,p,k,\alpha,R^2_\Omega)=1-(F^\alpha_{p-k,n-p-1}\times(p-k))/(n-p-1)\times(1/R^2_\Omega-1). \tag{5}$$

The value (5) in Theorem 2 is a data-dependent cut off point of the test for testing H_0 based on the relative R squared statistic R^2_ω/R^2_Ω. Therefore, the value (5) is suggested to be a data-dependent threshold of the relative R squared procedure.

Combining the above results, we propose a procedure of the RRSM by implementing a data-dependent threshold.

Acknowledgments

We thank for the reviewer's value comments. This work was partially supported by the National Science Council and National Center for Theoretical Sciences, Taiwan.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2013.08.002.

References

Ambros, V., 2004. The functions of animal microRNAs. Nature 431, 350–355.
Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297.
Bartonicek, N., Enright, A.J., 2010. SylArray: a web server for automated detection of mirna effects from expression data. Bioinformatics 26, 2900–2901.
Broderick, J.A., Zamore, P.D., 2011. MicroRNA therapeutics. Gene Therapy 18, 1104–1110.

Buse, A., 1973. Goodness of fit in generalized least squares estimation. The American Statistician 27, 106–108.
Cameron, A.C., Windmeijer, F.A.G., 1997. An *R*-squared measure of goodness of fit for dome common nonlinear regression models. Journal of Econometrics 77, 329–342.
Friedman, R.C., Farh, K.K., Burge, C.B., Bartel, D.P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Research 19, 92–105.
Gennarino, V.A., D′Angelo, G., Dharmalingam, G., Fernandez, S., Russolillo, G., Sanges, R., Mutarelli, M., Belcastro, V., Ballabio, A., Verde, P., Sardiello, M., Banfi, S., 2012. Identification of microRNA-regulated gene networks by expression analysis of target genes. Genome Research 22, 1163–1172.
Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P, Bartel, D.P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular Cell 27, 91–105.
Hsieh, W.J., Wang, H., 2011. Human MicroRNA target identification by RRSM. Journal of Theoretical Biology 286, 79–84.
Huang, J.C., Babak, T., Corson, T.W., Chua, G., Khan, S., Gallie, B.L., Hughes, T.R., Blencowe, B.J., Frey, B.J., Morris, Q.D., 2007a. Using expression profiling data to identify human MicroRNA targets. Nature Methods 4, 1045–1049.
Huang, J.C., Morris, Q.D., Frey, B.J., 2007b. Bayesian inference of MicroRNA targets from sequence and expression data. Journal of Computational Biology 14, 550–563.
Hsu, S.D., Lin, F.M., Wu, W.Y., Liang, C., Huang, W.C., Chan, W.L., Tsai, W.T., Chen, G.Z., Lee, C.J., Chiu, C.M., Chien, C.H., Wu, M.C., Huang, C.Y., Tsou, A.P., Huang, H.D., 2011. miRTarBase: a database curates experimentally validated MicroRNA-target interactions. Nucleic Acids Research 39, 163–169.
Le, T.D., Liu, L., Tsykin, A., Goodall, G.J., Liu, B., Sun, B.Y., Li, J., 2013. Inferring microRNA-mRNA causal regulatory relationships from expression data. Bioinformatics 29, 765–771.
Lewis, B.P., Burge, C.B., Bartel, D.P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120, 15–20.
Lu, I.L., Wang, H., 2012. Protein-specific scoring method for ligand discovery. Journal of Computational Biology 19, 1215–1226.
Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P., Hatzigeorgiou, A.G., 2009. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Research 37, 155–158.
Seber, George A.F., Lee, Alan J, 2003. Linear Regression Analysis, 2nd ed. John Wiley & Sons.
van Dongen, S., Abreu-Goodger, C., Enright, A.J., 2008. Detecting MicroRNA binding and siRNA off-target effects from expression data. Nature Methods 5, 1023–1025.
Wang, H., Li, W.-H., 2009. Increasing MicroRNA target prediction confidence by the relative R^2 method. Journal of Theoretical Biology 259, 793–798.
Wang, H., Wang, Y.-H., Wu, W.-S., 2011. Yeast cell cycle transcription factors identification by variable selection criteria. Gene 485, 172–176.