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