Journal of Theoretical Biology 360 (2014) 200–207

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Identifying local co-regulation relationships in gene expression data ☆

Yonggang Pei [a], Qinghui Gao [a,*], Juntao Li [a], Xiting Zhao [b]

[a] College of Mathematics and Information Science, Henan Normal University, Xinxiang 453007, China
[b] College of Life Science, Henan Normal University, Xinxiang 453007, China

ARTICLE INFO

Article history:
Received 15 December 2013
Accepted 26 June 2014
Available online 18 July 2014

Keywords:
Correlation
B-spline

ABSTRACT

Identifying interesting relationships between pairs of genes, presented over some of experimental conditions in gene expression data set, is useful for discovering novel functional gene interactions. In this paper, we introduce a new method for id entifying L ocal C o-regulation R elationships (IdLCR). These local relationships describe the behaviors of pairwise genes, which are either up- or down-regulated throughout the identified condition subset. IdLCR firstly detects the pairwise gene-gene relationships taking functional forms and the condition subsets by using a regression spline model. Then it measures the relationships using a penalized Pearson correlation and ranks the responding gene pairs by their scores. By this way, those relationships without clearly biological interpretations can be filtered out and the local co-regulation relationships can be obtained. In the simulation data sets, ten different functional relationships are embedded. Applying IdLCR to these data sets, the results show its ability to identify functional relationships and the condition subsets. For micro-array and RNA-seq gene expression data, IdLCR can identify novel biological relationships which are different from those uncovered by IFGR and MINE.

1. Introduction

For a data set with tens of thousands of variables, which may contain various, undiscovered relationships, how can we efficiently identify the interesting relationships? One way to begin exploring the data set is to determine what kinds of relationships we are interested in, then to search for variable pairs with these relationships using data mining. The methods following this way face two issues, defining similarity and robust discovery. The former is related to determining what kinds of relationships we are interested in. The latter concerns how to remove noises as much as possible. These methods can be divided into the following two categories, biclustering methods and the others.

In biclustering methods, a submatrix of a data set is identified by using various definitions of similarity for interesting relationships. Variables express the similar behavior over columns in the submatrix. The definitions of similarity employed in these methods are based on the constant model (Hartigan, 1975; Busygin et al., 2002), the additive model (Getz et al., 2000; Califano et al., 2004; Sheng et al., 2003), the multiplicative model (Cheng and Church, 2000; Yang et al., 2003; Wang et al., 2002; Lazzeroni and Owen, 2002; Segal et al., 2003; Tang et al., 2001; Klugar et al., 2003; Hochreiter et al., 2010), the general linear model (Gan et al., 2005, 2008; Zhao et al., 2008), and correlation (Mosca et al., 2009; Bhattacharya and De, 2009; Mosca et al., 2009; Nepomuceno et al., 2011; Gao et al., 2012). These biclustering methods can delete the redundant samples and obtain some robust discoveries. However, because of the restriction of those definitions of similarity, these methods can only capture the linear relationships, which are some of the interesting relationships describing the similar behaviors by genes in gene expression data.

In the other category, the pairs are ranked by their scores which are calculated according to some similarity measure of dependence for the pair. Then the top-scoring pairs should be examined. These methods include IFGR (Mosca et al., 2009) and MINE (Reshef et al., 2011), in which the top-scoring pairs are the robust discoveries. IFGR employs a definition of similarity based on correlation and genetic algorithm to delete the irrelevant samples, and then ranks the pairs to get the robust discoveries. But it also captures the linear relationships because of the restriction of the definition of similarity. MINE employs a maximal information coefficient to identify a wide range of relationships. However, since the method does not consider that the irrelevant or redundant samples mingling in data set, some pairs of variables expressing their close associations over some of the samples get

☆This work is supported by National Natural Science Foundation of China (61203293, 31372105), Program for Science and Technology Innovation Talents in Universities of Henan Province (13HASTIT040), Foundation of Henan Educational Committee (13A120524), Henan Higher School Funding Scheme for Young Teachers (2012GGJS-063) and PhD research funds of Henan Normal University (QD13041).

* Corresponding author.
E-mail addresses: peiyonggang@htu.edu.cn (Y. Pei), qinghuigaocb@163.com (Q. Gao), juntaolimail@126.com (J. Li), zhaoxt0411@126.com (X. Zhao).

http://dx.doi.org/10.1016/j.jtbi.2014.06.032
0022-5193/

low scores. Then, this kind of interesting relationships over some sample subset is missed.

In this paper, we propose a new method for identifying local co-regulation relationships (IdLCR). IdLCR firstly detect the associations taking various functional forms by using a regression spline model. However, many of the functional relationships are not biologically meaningful. We are only interested in those functional relationships which can describe the close associations between pairs of variables and the positive and negative regulations. Then, by employing an appropriately punished Pearson correlation, IdLCR can select the local co-regulation relationships from the detected functional ones because Pearson correlation itself is a measure of linear dependence, which are local co-regulated. These identified relationships can describe both the close associations between pairs of variables and the similar behaviors expressed by them.

For simulation data sets, the results indicate that IdLCR can identify various embedded functional relationships only by using the regression spline model. For both microarray and RNA-seq gene expression data, the results from running IdLCR show that the novel biological relationships are identified. Our method is also comparable with, if not better than, IFGR and MINE in identifying different biological relationships. Source code for IdLCR is available upon request.

2. Methods

We assume that there are some different pairs of variables (x and y), and for each pair, there are n observation pairs. Let $N=\{1,2,\ldots,n\}$ be the index set. We firstly aim to detect the functional relationships fitted by $|I|$ observation pairs of x and y, where $I\subset N$ and $|I|$ is the size of the index subset I. Then, we identify the co-regulation relationships, which describe the close associations between pairs of variables, from all the detected functional relationships. Hence, we split discussion of our methodology into two parts. In Section 2.1, we propose a regression spline model and give the discussions for the model. In Section 2.2, we motivate and introduce a punished Pearson correlation.

2.1. Regression spline model

2.1.1. The model

Denote by m the functional relationship between x and y. Because of the irrelevant or redundant samples mingling in, only $|I|$ observation pairs of x and y satisfy $y_i=m(x_i)+\epsilon_i$ $(i\in I)$, where each ϵ_i is the observation of a random variable denoting the variation of y around $m(x)$. Since m is a smooth function from $\mathbb{R}$ (the set of the real number) to $\mathbb{R}$, the following regression spline model y_i can be constructed to approximate it:

$$y_i=\sum_{k=1}^{K}\gamma_k b_k(x_i)+\epsilon_i=B(x_i)^T\Gamma+\epsilon_i,\quad i\in I, \tag{1}$$

where $\{b_1,\ldots,b_K\}$ is a prescribed set of uniform B-spline basis functions with the degrees d, the coefficients $\gamma_1,\ldots,\gamma_K$ are called control points or de Boor points and K is the number of control points. The second equality above holds if B is the basis function vector containing the B-spline basis functions $b_k, k=1,\ldots,K$ as elements and $\Gamma=(\gamma_1,\ldots,\gamma_K)^T$.

In the model (1), the spline curve can be determined when Γ is known. In order to estimate the parameter vector Γ, we assume that, for the n observations $y_1,\ldots,y_n$, $y_i\sim N(B(x_i)\Gamma,\sigma^2), i\in I$ and $y_j\sim N(\mu_y,\sigma_y^2), j\in J=N\backslash I$. Then the joint density function for the $|N|$ observations can be written as the following form:

$$f(y|\Gamma,\sigma^2,\mu_y,\sigma_y^2)=\prod_{i\in I}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y_i-B(x_i)^T\Gamma)^2}{2\sigma^2}\right)\prod_{j\in J}\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(\frac{-(y_j-\mu_y)^2}{2\sigma_y^2}\right). \tag{2}$$

The form (2) is a likelihood for evaluating how well the points fit the model. The higher the likelihood is, the better the points fit the model. By maximizing the likelihood, we get the following maximum likelihood estimators (MLEs):

$$\widehat{\Gamma}=W\left(\sum_{i\in I}y_i B(x_i)\right), \tag{3}$$

$$\widehat{\sigma^2}=\frac{1}{|I|}\sum_{i\in I}(y_i-B^T(x_i)\widehat{\Gamma})^2, \tag{4}$$

$$\widehat{\mu_y}=\frac{1}{|J|}\sum_{j\in J}y_j, \tag{5}$$

$$\widehat{\sigma_y^2}=\frac{1}{|J|}\sum_{j\in J}(y_j-\widehat{\mu_y})^2, \tag{6}$$

where $W=(\sum_{i\in I}B(x_i)B^T(x_i))^{-1}$ if $\sum_{i\in I}B(x_i)B^T(x_i)$ is invertible, otherwise, W is the pseudo inverse matrix of $\sum_{i\in I}B(x_i)B^T(x_i)$.

Upon constructing the MLEs, we can identify the functional relationship if the index subset I is determined. However, the two tasks, identifying the functional relationship, determining I, are coupled each other. And when one is finished, the other is easy to get. Here, we employ EM algorithm to deal with that.

The algorithm is outlined as below:

Step 1. All indices of the observations are assigned randomly to the index subset I and $J=N\backslash I$, estimate the parameters: ${}^{(0)}\Gamma$, ${}^{(0)}\sigma^2$, ${}^{(0)}\mu_y$, ${}^{(0)}\sigma_y^2$.

Step 2. In the $(k+1)$th iteration, assign each index s to the index subset with minimum deviation. The deviation for the index subset I is measured by

$$L_{s,I}^{(k)}=-\log\left(\prod_{i\in I}\frac{1}{\sqrt{2\pi({}^{(k)}\sigma^2)}}\exp\left(\frac{-(y_i-B^T(x_i)({}^{(k)}\Gamma))^2}{2({}^{(k)}\sigma^2)}\right)\right).$$

The deviation for the index subset J is

$$L_{s,J}^{(k)}=-\log\left(\prod_{j\in J}\frac{1}{\sqrt{2\pi({}^{(k)}\sigma_y^2)}}\exp\left(-(y_j-\frac{({}^{(k)}\mu_y)^2}{2({}^{(k)}\sigma_y^2)})\right)\right).$$

Step 3. Compute ${}^{(k)}\Gamma$, ${}^{(k)}\sigma^2$, ${}^{(k)}\mu_y$, ${}^{(k)}\sigma_y^2$.

Step 4. Repeat 2 till convergence.

2.1.2. Discussions

Note that the determined functional relationship is affected by the choices of the number of control points K and the degree of the spline basis functions d. For different K and d, the constructed regression spline models are different, then the determined functional relationships are also different. Which one is the right one? We give our choices and grounds.

In a B-spline, the degree d controls the smoothness of a spline curve. The bigger the degree is, the more smooth the spline cure is. However, with increasing of the degree, blending function is more difficultly to precalculate. When the degree d is equal to 3, the spline cure is begin to be smooth and the blending function is also easily to precalculate. Hence, cubic B-spline with uniform knot-vector is the