Advances in Applied Mathematics 应用数学进展, 2023, 12(5), 2177-2186
Published Online May 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.125222

Hans汉斯

基于回归理论恢复基因调控网络

张　雪，严传魁*

温州大学数理学院，浙江　温州

收稿日期：2023年4月17日；录用日期：2023年5月9日；发布日期：2023年5月17日

摘　要

基因之间的调控关系隐含在基因表达数据里，需要分析该数据从而揭示基因调控网络的拓扑结构。由于静态基因表达数据的样本较少，因此本文提出基于距离相关性扩充样本数据量的方法。接着，本文提出恢复基因调控网络拓扑结构的方法，基于距离样本数据根据回归理论建立基因调控网络线性回归模型，对模型应用最小二乘估计和假设检验判断基因之间是否存在调控关系。此外，提出可以控制假阳性的方法，利用统计检验控制错误发现率提高模型预测的准确性。最后，在DREAM3数据集上验证方法的可行性。

关键词

基因调控网络，最小二乘估计，假设检验

Restoration of Gene Regulatory Network Based on Regression Theory

Xue Zhang, Chuankui Yan*

College of Mathematics and Physics, Wenzhou University, Wenzhou Zhejiang

Received: Apr. 17th, 2023; accepted: May 9th, 2023; published: May 17th, 2023

Abstract

The regulatory relationship between genes is implicit in the gene expression data, which needs to be analyzed to reveal the topology of the gene regulatory network. Since the small sample size of static gene expression data, this paper proposes a method to expand the sample data size based on distance correlation. Then, this paper proposes a method to restore the topology of gene regulatory network. Based on the distance sample data, a linear regression model of gene regulatory net-

*通讯作者。

文章引用: 张雪, 严传魁. 基于回归理论恢复基因调控网络[J]. 应用数学进展, 2023, 12(5): 2177-2186.
DOI: 10.12677/aam.2023.125222

work is established according to the regression theory. The least square estimation and hypothesis testing are applied to the model to determine whether there is a regulatory relationship between genes. In addition, a method to control false positives is proposed. Statistical test is used to control the false discovery rate to improve the accuracy of model prediction. Finally, the feasibility of the method is verified on the DREAM3 dataset.

Keywords

Gene Regulatory Network, Least Square Estimation, Hypothesis Testing

1. 引言

基因调控网络(Gene Regulatory Network, GRN)是由细胞内基因之间相互作用构成的网络。节点代表基因，节点间的连边代表基因之间的调控关系。它是一个抽象的概念，但可以利用数学表达式来建立基因调控的过程，从而帮助人们理解细胞内基因之间的关系。重构基因调控网络属于逆向工程问题，目前该领域的学者提出了很多用于研究基因调控网络的数学模型[1]。例如最简单的、离散的布尔网络(Boolean Networks)模型[2]；一种概率描述的贝叶斯网络(Bayesian Networks)模型[3]以及微分方程(Differential Equation)模型[4]等。

基因表达数据反映的是直接或间接测量得到的基因转录产物 mRNA 在细胞中的丰度，它包含的基因活动信息能够反映当前细胞的状态，采用反向工程的方法分析基因表达数据可以还原基因的调控网络。基因表达数据可以分为时间序列基因表达数据和静态基因表达数据，它们都是基因表达水平的数据矩阵。时间序列基因表达数据记录了生命活动周期或不同状态下连续变化的基因表达水平，反映了基因调控网络的动态变化，可以通过相应的算法推断基因之间的调控关系[5]；静态基因表达数据记录了稳定状态下细胞内的基因表达水平，包含了基因间的调控信息，直接反映了基因之间的调控关系，可以通过基于信息论等方法建立基因调控网络[6] [7]。静态基因表达数据具有高维度、小样本的特点，即与大量的基因相比，样本数量较少。对于这一困难多种方法被提出验证。Fujita 等对于样本数据量低于基因数量的问题，提出 SVAR 方法建模基因表达调控网络[8]；Barrera 等利用互信息降维的方法，通过计算基因之间的互信息，最终选择基因之间互信息较高的基因恢复基因调控网络[9]；Székely 等扩充了样本数据量，通过距离相关性概念将 m 个样本数据扩充为 m^2 个样本数据。通过增加样本数据[10]，可以提高模型的预测性能。

真实生物体基因间的调控关系错综复杂，模型识别真正的边显得十分重要。而模型在恢复基因调控网络拓扑结构过程中可能会出现真实基因中没有调控关系但被模型错误的预测为有调控关系的情况，即预测结果中出现比较多假阳性边的情况。因此需要降低假阳性边的数量，提高模型预测的准确率。Fujita 等提出通过统计检验控制错误发现率的方法[8]。罗霄提出通过 DSWLasso 逐步删除恢复的基因调控网络中的弱调控关系的方法[11]。

本文在此基础上进行研究，将基因调控网络的恢复问题转化为识别方程参数的问题。通过假设检验判断基因间是否存在调控关系，此时将基因调控网络拓扑结构恢复问题转化为分类问题进行解决。并采取适当的方法减少结果中出现的假阳性边的数量。

2. 方法

回归分析是用来确定两种或两种以上变量之间相互依赖的定量关系的一种统计方法。本文所研究的基因调控网络包含多个自变量，并且因变量和自变量之间呈线性关系，因此可以利用多元线性回归模型判断基因调控网络中基因间的调控关系。

2.1. 基因调控网络线性回归模型

基于距离样本数据采用回归理论建立基因调控网络线性回归模型，以此判断基因 i 与其他基因之间是否存在调控关系，模型表示为：

$$X_i = X_{\setminus i} B_i + \varepsilon_i \tag{1}$$

其中 $X_{\setminus i} = \left(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_N\right)$ 表示除第 i 个基因外，其余 $N-1$ 个基因的距离样本向量，ε_i 代表误差，满足相互独立且分布服从 $N\left(0, \sigma^2\right)$。

样本数据代入(1)中，则(1)可以改写为：

$$Y_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iM} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{i-1,1} & x_{i+1,1} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{i-1,2} & x_{i+1,2} & \cdots & x_{N2} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ x_{1M} & x_{2M} & \cdots & x_{i-1,M} & x_{i+1,M} & \cdots & x_{NM} \end{pmatrix} \begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{i,i-1} \\ b_{i,i+1} \\ \vdots \\ b_{iN} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{i,i-1} \\ \varepsilon_{i,i+1} \\ \vdots \\ \varepsilon_{iN} \end{pmatrix} = XB_i + E_i \tag{2}$$

恢复基因调控网络拓扑重构解释为从可测变量 X 和 Y_i 中估计 B_i，则转化为求解多元线性回归方程(2)的参数 B_i 的问题。在这里采用对每个基因进行单独回归分析的方法，用最小化误差的平方和估计每个基因的 B_i，得到预测的邻接矩阵 $\hat{B}$。损失函数利用均方误差的形式，同时它也是目标函数：

$$L\left(B_i\right) = \frac{1}{2}\left(XB_i - Y_i\right)^{\mathrm{T}}\left(XB_i - Y_i\right) \tag{3}$$

为确定模型的参数 B_i，同时使预测值 $\hat{Y}_i$ 尽可能地接近真实值 Y_i，等价于求 $L\left(B_i\right)$ 的最小值。由于目标函数(3)中包含要求解的参数 B_i，因此在求解(3)最小值的过程中便可以将模型(2)中的参数 B_i 求解出来，在这里我们选取最小二乘法求解。用最小二乘法估计 B_i，只需要令其对 B_i 的一阶偏导数为 0，偏导数为 0 的点是(3)的最小值点，求得：

$$\hat{B}_i = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y_i,\ i = 1, \cdots, N \tag{4}$$

(4)就是 B_i 的最小二乘估计，该方法适用于 $X^{\mathrm{T}}X$ 的逆矩阵存在 $\left(M > N\right)$ 时的情况。此时 $\hat{B}_i$ 是 B_i 的无偏估计，无偏性保证了最小二乘估计具有较好的结果。

2.2. 基因表达数据处理

基因之间的调控关系隐含在基因表达数据里，需要分析该数据从而揭示基因调控网络的拓扑结构。而现实难以收集到大量的基因表达数据，但对于(4)式，当 $M < N$ 时，$X^{\mathrm{T}}X$ 不可逆，此时最小二乘估计失效，因此需要扩充样本量。本文根据 Székely 的方法引入距离相关性来扩充样本量。假设有 N 个基因，实验收集到这 N 个基因在 m 个样本上的表达水平，记为 $G = \left(G_1, G_2, \cdots, G_N\right)$，其中 $G_i = \left(g_{i1}, g_{i2}, \cdots, g_{im}\right)^{\mathrm{T}}$，$i = 1, \cdots, N$。根据距离相关性和 G 可以生成一个矩阵 J，J 包含 m^2 个元素，各元素

定义为：

$$j_{sp}=\left\|g_{is}-g_{ip}\right\|_2, i=1,2,\cdots,N$$

其中$\|\cdot\|_2$表示l_2范数。对矩阵J中心化可得对称矩阵D，D中元素定义为：

$$d_{sp}=j_{sp}-\frac{1}{m}\sum_{k=1}^{m}j_{sk}-\frac{1}{m}\sum_{k=1}^{m}j_{kp}+\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}j_{kl}$$

将D中下三角元素重新组合为一列，可得到一个包含$M=\frac{m(m+1)}{2}$个元素的列向量X_i，X_i就是对原始基因表达数据G_i的扩充，由原来的m个样本扩充为M个样本，因此达到了扩充样本基因表达数据的目的，称$X=(X_1,X_2,\cdots,X_N)$为距离样本数据。

2.3. 显著性检验

2.3.1. 基因调控网络线性回归模型的可行性

本文采用多元线性回归方程预测基因i和其它基因间的调控关系，建立的恢复模型为(2)。但基因i和其它基因之间是否为线性关系是未知的，因此需要判断线性关系是否成立。通过对模型(2)进行整体检验来判断回归系数是否不全为0，如此便可说明其他基因是否对基因i有显著影响。如果所有回归系数为0，则代表它们对基因i没有线性调控关系，因此这相当于检验原假设H_0^i是否成立：

$$H_0^i: b_{i1}=b_{i2}=\cdots=b_{i,i-1}=b_{i,i+1}=\cdots=b_{iN}=0,\ i=1,\cdots,N$$

H_0^i对应的备择假设为：

$$H_1^i: b_{ij}\neq 0,\ j=1,\cdots,N-1$$

如果有充分的理由说明H_0^i不成立，那么意味着基因i与其它$N-1$个基因之间的线性关系成立。对模型整体检验选取F检验。在H_0^i为真时构造检验统计量：

$$S_R=\sum\left(X_{\setminus i}\hat{B}_i-\overline{Y}_i\right)^2$$

$$S_e=\sum\left(Y_i-X_{\setminus i}\hat{B}_{ii}\right)^2$$

$$F=\frac{S_R/(N-1)}{S_e/(M-N-2)}\sim F(N-1,M-N-1)$$

其中$F(N-1,M-N-1)$表示自由度为$(N-1,M-N-1)$的F分布，取显著性水平为$\alpha(0<\alpha<1)$。我们选定$\alpha=0.05$，当$F\geq F_{1-\alpha}(N-1,M-N-1)$时拒绝原假设$H_0^i$，认为基因$i$与其它$N-1$个基因之间有线性关系，说明利用基因调控网络线性回归模型恢复拓扑结构是可行的。

2.3.2. 基因调控关系的判定

模型整体检验说明基因i与其它$N-1$个基因之间的线性关系成立，但是不能确定某个基因j对基因i是否有调控关系。因此需要对回归系数进行检验，即讨论基因与基因之间是否有调控关系。

本节采用的基因调控矩阵元素为0和1，其中0代表两个基因之间没有调控关系，1表示两个基因之间有调控关系。利用模型(2)判断哪些基因之间存在调控关系，等价于寻找非0的b_{ij}。因此为检验基因i与其他基因之间是否存在调控关系，建立如下假设：

$$H_0^{ij}: b_{ij}=0,\ H_1^{ij}: b_{ij}=1$$

如果有充分的理由说明 H_0^{ij} 不成立，那么就可以接受 H_1^{ij}，认为基因 i 与基因 j 之间存在调控关系。选取的假设检验方法为 t 检验。已知 $\hat{b}_{ij} \sim N\left(b_{ij}, u_{jj}\sigma^2\right)$，当假设 H_0^{ij} 为真时构造检验统计量：

$$t_{ij} = \frac{\hat{b}_{ij}}{\sqrt{u_{jj}\hat{\sigma}^2}} \sim t(m-N-1)$$

其中 $t(m-N-1)$ 表示自由度为 $(M-N-1)$ 的 t 分布，u_{jj} 为 $\left(X^{\mathrm{T}}X\right)^{-1}$ 的第 j 个对角线元素，$\hat{\sigma}^2 = \frac{S_e}{M-N-1}$ 为 σ^2 的无偏估计。取显著性水平为 $\alpha(0<\alpha<1)$。我们选定 $\alpha = 0.05$，当 $|t_{ij}| \geq t_{1-\frac{\alpha}{2}}(M-N-1)$ 时拒绝原假设 H_0^{ij}，接受备择假设 H_1^{ij}，认为 $b_{ij}=1$，说明基因 j 对基因 i 的作用显著，即确定基因 j 对基因 i 有调控关系；反之，当 $|t_{ij}| < t_{1-\frac{\alpha}{2}}(M-N-1)$ 时不拒绝原假设 H_0^{ij}，则认为 $b_{ij}=0$，说明基因 j 对基因 i 的作用不显著，即确定基因 j 对基因 i 没有调控关系。

通过对模型(2)基于距离样本数据利用最小二乘估计得到初始的预测矩阵后，接着对回归系数进行假设检验便可以确定 $\hat{B}$ 中元素为 1 的位置，$\hat{B}$ 中元素为 1 意味着基因之间有调控关系，此时模型只保留了对基因 i 有显著作用的调控基因。因此可根据构建的模型确定具有调控关系的边，即完成了恢复基因调控网络的拓扑结构。

对模型恢复的基因调控网络拓扑结构进行整体评估，定义评价指标：

$E = \frac{FN}{FN+TP}$ 为模型未识别率，其中 FN 表示基因间有调控关系而模型没有识别出的个数，$FN+TP$ 表示基因间真实又调控关系的个数；

$FPR = \frac{FP}{FP+TN}$ 为假阳性率，其中 FP 表示基因间没有调控关系而模型识别为有调控关系的个数，$FP+TN$ 表示基因间没有调控关系的个数；

$FDR = \frac{FP}{FP+TP}$ 为误报率，其中 $FP+TP$ 表示模型识别基因间有调控关系的个数。

2.4. 控制假阳性结果的数量

t 检验的拒绝域为 $W = \left\{|t| \geq t_{1-\frac{\alpha}{2}}(M-N-1)\right\}$，而由于随机性等因素的干扰，在进行判断时可能出现两种错误，假阴性错误和假阳性错误。若出现原假设成立，但计算得到的 b_{ij} 落入拒绝域 W 中的错误，称为假阳性错误(第I类错误)。在 Neyman-Pearson 假设检验中，可以要求假阳性率在可接受的范围内，设置的检验水平 $\alpha = 0.05$ 就是限制犯假阳性错误的概率，通过 α 达到对单重假设检验的统计推断的错误控制。在单个假设检验中假阳性率是可控的，但是对于整体多个假设而言，假阳性结果会随着检验次数的增加而增大，出现超出预设范围的情况。因此需要对回归系数整体进行多重检验校正，减少出现假阳性结果的次数。选择控制 FDR 的 BH(Benjamini-Hochberg)法，该方法可以将假阳性率控制在统计显著水平 α 以下。调整 p 值控制 FDR，过程如下：

对于假设检验 $\left\{H_0^{ij}\right\}$ $i, j = 1,\cdots,N, i \neq j$，$H_0^{ij}$ 表示第 ij 个检验的原假设。将所有检验的 p 值从小到大排序生成顺序数 $\{p_k\}$, $k = 1,\cdots,v$。令 $q_{\max} = p_{\max}$，定义

$$q_k = \min\left\{p_k \frac{v}{k}, p_{k-1}\right\},\ k = 1,\cdots,v-1$$

其中 v 代表检验次数。当 $q_k < \alpha$ 时，拒绝前 k 个假设，认为基因之间有调控关系。

3. 模型应用及恢复结果

3.1. DREAM 数据集上的实验

通过以上分析可以根据模型(2)推断基因调控网络，为了验证其可行性，本节在 DREAM(Dialogue for Reverse Engineering Assessments and Method)提供的静态基因表达数据上进行实验。采用 DREAM3 数据集，数据集可在 https://gnw.sourceforge.net/dreamchallenge.html 下载。DREAM3 数据集提供了静态基因表达数据和确定的基因调控网络，本节实验选取数据集 InSilicoSize10-Ecoli1-null-mutants，它包含 10 个基因，将基因标号为 $G1,\cdots,G10$。该数据集对应的网络结构如图 1 所示。其中红线代表基因之间是抑制作用，蓝线代表基因之间是激活作用。本文不区分基因之间的抑制和激活作用，主要关注基因之间是否存在调控关系。

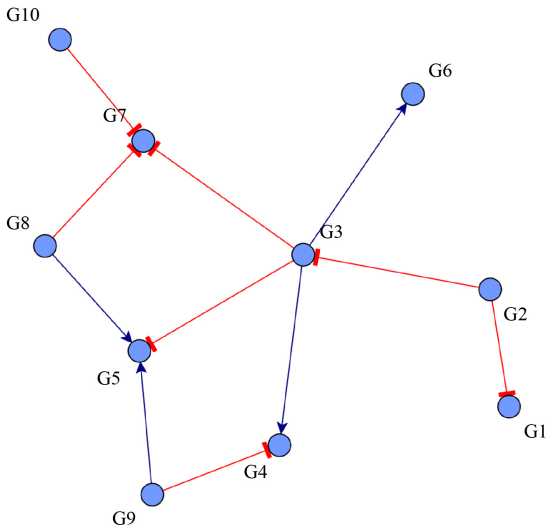

Figure 1. Topology of gene regulatory network
图 1. 基因调控网络拓扑结构

由于实验会受到随机性的影响，为了确保模型预测的稳定性，结果为多次实验的平均。每次实验对距离样本采用随机抽样用于模型(2)来构建基因调控网络，每次抽样循环多次并取平均值作为最终结果。

3.2. 实验结果及讨论

3.2.1. 基因调控网络线性回归模型的可行性

DREAM3 提供了确定的基因调控网络拓扑图，本文只考虑基因之间是否存在调控关系，因此调控矩阵中的元素只有 0，1 两种。数据集 InSilicoSize10-Ecoli1-null-mutants 中仅包含 11 个基因表达数据，首先对其进行扩充得到包含 66 个样本数据距离样本，从中随机选取不同长度的样本用于恢复拓扑结构。

对每个基因在不同样本数据时的模型(2)进行检验以判断利用其恢复基因调控网络拓扑结构是否可行，表 1 包含了在不同数据下对每一个模型进行整体检验的 F 值。

在 $\alpha=0.05$ 情况下，由表 1 以及 Friedman 检验表可知，对于 G8，G9，G10 基因，由于 $F<F_{1-\alpha}(9,M-11)$，因此不拒绝原假设 H_0^i, $i=8,9,10$，认为回归方程的所有系数为 0，与真实的基因调控网络一致；对于另外 7 个基因，由于 $F\geq F_{1-\alpha}(9,M-11)$，因此拒绝原假设 H_0^i，认为每一个基因 $i,(i=1,\cdots,7)$ 与其它基因之间有线性关系，说明利用线性回归模型预测基因调控矩阵是可行的。

Table 1. F-value of the overall model test
表 1. 模型整体检验的 F 值

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
M = 15	49.01	228.71	106.3	3.84	6.19	36.16	7.29	3.11	5.16	2.86
M = 25	93.96	116.11	107.96	3.46	3.49	22.01	3.14	1.40	2.14	1.76
M = 45	112.40	128.70	152.82	3.72	3.02	31.03	4.28	1.39	1.86	2.16

3.2.2. 基因调控关系的判定及控制假阳性结果的数量

对于原假设 $H_0^{ij}:b_{ij}=0$ 和备择假设 $H_1^{ij}:b_{ij}=1$，采用 t 检验可以判定矩阵 $\hat{B}$ 中位置为 1 的元素，即可以判定基因之间的调控关系。

图 2 显示曲线呈下降趋势，模型的未识别率 E 在逐渐降低，说明随着数据量 M 的增加，模型识别到的连接个数逐渐增加。但初始样本数据会限制模型对基因调控网络拓扑结构的恢复能力，当 $M=55$ 时，$E=0.4802$，此后距离样本再增加不会明显提高模型的恢复能力。

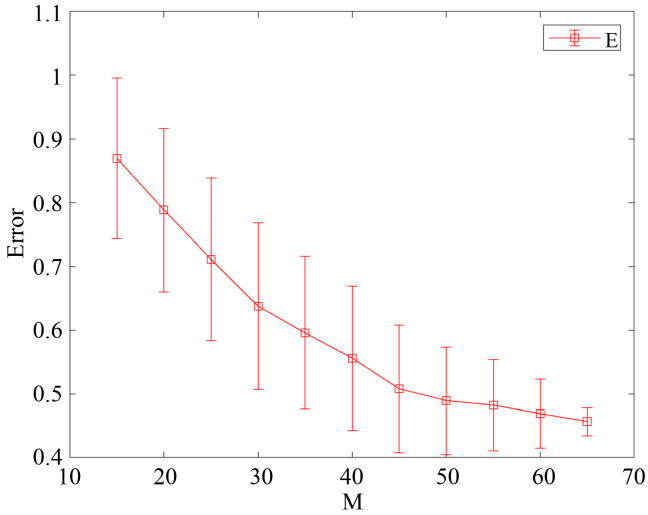

Figure 2. For the gene regulatory network linear regression model (2), the unrecognized rate E with its standard deviation under different sample data size M
图 2. 对于基因调控网络线性回归模型(2)，不同样本数据量 M 下的未识别率 E 及其标准差

由图 3 可知，$FP=5$，说明模型(2)在恢复基因间的调控关系时出现较多的假阳性边，即出现较多实际基因没有调控关系但模型错误的判断为有调控关系的结果。图 4 是在 $M=40$ 时对模型(2)应用控制 FDR 的 BH 法后恢复的基因调控网络中假阳性边的情况。由图可知 $FP=1$，此时假阳性率低于 0.05。与图 3 相比，模型预测的假阳性由 5 条减少为 1 条，达到了控制模型识别假阳性边的个数的目的。因此对于在构造基因调控网络出现假阳性边较多的情况，可以采用控制 FDR 的 BH 法来减少出现模型恢复过程中出现假阳性边的次数。

不失一般性，在不同数据量时均可对整体多个假设采用 FDR 校正。由图 5 可得，对于不同的样本数据量 M，FDR 校正后假阳性率较之前会降低，达到了对构建的基因调控网络假阳性边数量控制的目的，假阳性率始终在(0, 0.05)的范围之内。

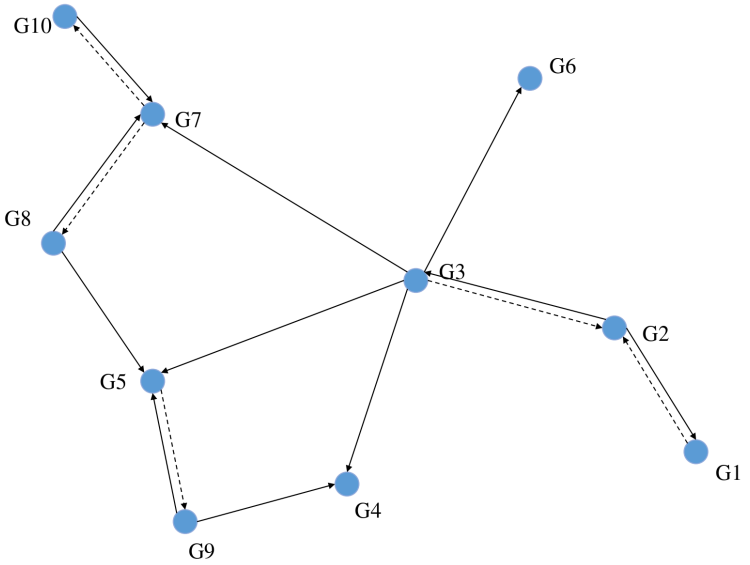

Figure 3. For the gene regulatory network linear regression model (2), the restored gene regulatory relationship when $M = 40$. The solid line represents the true gene regulatory relationship, and the dashed line indicates the edges where there is no regulatory relationship between genes but the model identifies as having a regulatory relationship

图 3. 对于基因调控网络线性回归模型(2)，$M = 40$ 时模型恢复的基因调控关系。实线表示真实的基因调控关系，虚线表示基因间不存在调控关系但模型识别为有调控关系的边

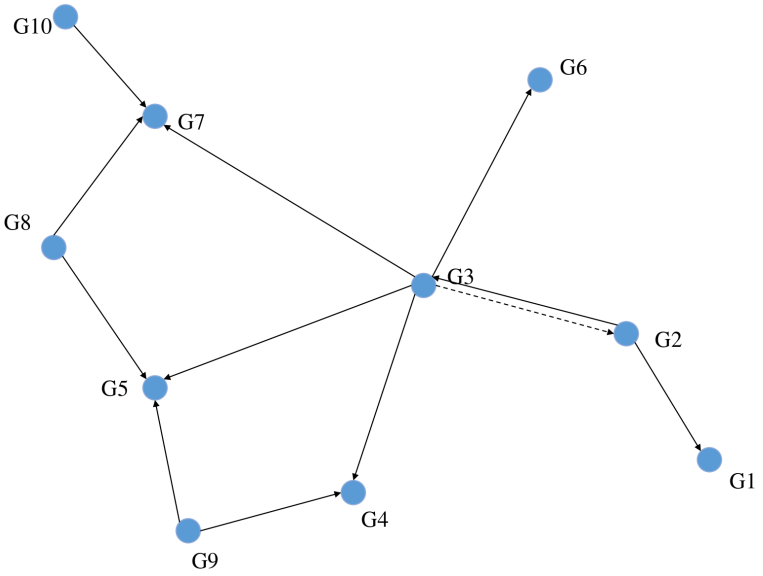

Figure 4. For the gene regulatory network linear regression model (2), the restored gene regulatory relationship when $M = 40$ after FDR correction. The solid line represents the true gene regulatory relationship, and the dashed line indicates the edges where there is no regulatory relationship between genes but the model identifies as having a regulatory relationship

图 4. 对于基因调控网络线性回归模型(2)，经过 FDR 校正后在 $M = 40$ 时模型恢复的基因调控关系。实线表示真实的基因调控关系，虚线表示基因间不存在调控关系但模型识别为有调控关系的边

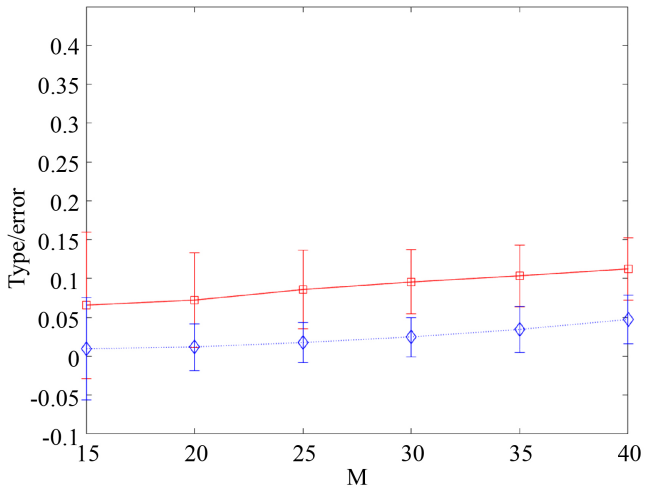

Figure 5. For the gene regulatory network linear regression model (2), the probability and standard deviation of false positive error under different sample data size M. The red line and blue line respectively represent FPR values before and after FDR correction
图 5. 对于基因调控网络线性回归模型(2)，不同样本数据量 M 下犯假阳性错误的概率及其标准差。红线蓝线分别代表 FDR 校正前和校正后的 FPR 的值

由于假阳性错误和假阴性错误在相同的样本数据量下不能同时增大或减小，只有增加样本数据量才可以使两种错误同时减少，因此采用 FDR 校正降低模型识别假阳边的数量后，不可避免的会增加假阴性边的个数。但由图 6 可知随着数据量的增加，FDR 校正前后模型出现假阴性错误的差距越来越小。

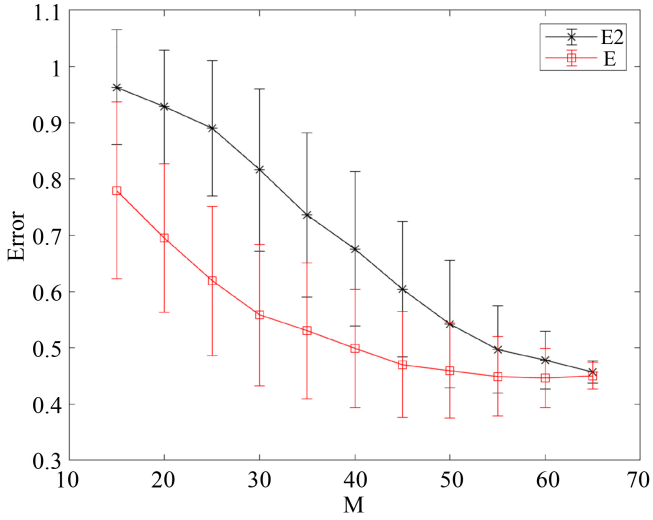

Figure 6. For the gene regulatory network linear regression model (2), the unrecognized rate E with its standard deviation under different sample data size M. The red and black lines represent the unrecognized rate E before and after correction, respectively
图 6. 对于基因调控网络线性回归模型(2)，不同样本数据量 M 下的未识别率 E。红色线和黑色线分别代表校正前和校正后的未识别率 E

4. 结论

基因调控网络存在于生物体内，是由细胞内基因之间相互作用构成的网络，可以利用数学表达式建立基因调控的过程，这可以帮助我们了解细胞内基因之间的关系。静态基因表达数据记录了稳定状态下细胞内的基因表达水平，可以反映基因之间的调控关系，但是其样本数较少，本文对此介绍了扩充样本数据产生距离样本的方法，并基于距离样本根据回归理论建立了基因调控网络线性回归模型，通过最小二乘估计和假设检验来确定哪些基因之间存在调控关系。接着应用控制 FDR 的 BH 法有效地减少了模型识别假阳性结果的数量，提高模型预测的准确率。最后选取 DREAM3 数据集验证模型的可行性，并在 $M = 40$ 时直观展现了该方法降低假阳性边的有效性。

基金项目

本文工作由国家自然科学基金(No. 11502062)支持。

参考文献

[1] Bansal, A.K., *et al*. (2005) The Role of Reverse Engineering in the Development of Generic Formulations. *Pharmaceutical Technology*, **29**, 50-55.

[2] Kauffman, S. (1969) Homeostasis and Differentiation in Random Genetic Control Networks. *Nature*, **224**, 177-178. https://doi.org/10.1038/224177a0

[3] Friedman, N. (2004) GraPhieal: Inferring Cellular Networks Using Probabilistie Models. *Science*, **303**, 799-805. https://doi.org/10.1126/science.1094068

[4] Gennemark, P. and Wedelin, D. (2007) Efficient Algorithms for Ordinary Differential Equation Model Identification of Biological Systems. *IET Systems Biology*, **1**, 120-129. https://doi.org/10.1049/iet-syb:20050098

[5] Rubiolo, M., Milone, D.H. and Stegmayer, G. (2015) Mining Gene Regulatory Networks by Neural Modeling of Expression Time-Series. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, **12**, 1365-1373. https://doi.org/10.1109/TCBB.2015.2420551

[6] Luo, W., Hankenson, K.D. and Woolf, P.J. (2008) Learning Transcriptional Regulatory Networks from High Throughput Gene Expression Data Using Continuous Three-Way Mutual Information. *BMC Bioinformatics*, **9**, 467. https://doi.org/10.1186/1471-2105-9-467

[7] Krämer, N., Schäfer, J. and Boulesteix, A.L. (2009) Regularized Estimation of Large-Scale Gene Association Networks Using Graphical Gaussian Models. *BMC Bioinformatics*, **10**, 384. https://doi.org/10.1186/1471-2105-10-384

[8] Fujita, A., Sato, J.R., Garay-Malpartida, H.M., *et al*. (2007) Modeling Gene Expression Regulatory Networks with the Sparse Vector Autoregressive Model. *BMC Systems Biology*, **1**, 1-11. https://doi.org/10.1186/1752-0509-1-39

[9] Barrera, J., Jr, R., Jr, D., *et al*. (2004) A New Annotation Tool for Malaria Based on Inference of Probabilistic Genetic Networks. *Critical Assessment of Microarray Data Analysis*, 36-40.

[10] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and Testing Dependence by Correlation of Distances. *Annals of Statistics*, **35**, 2769-2794. https://doi.org/10.1214/009053607000000505

[11] 罗霄. 基因调控网络构建方法研究[D] : [硕士学位论文]. 大连: 大连理工大学, 2020.