PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2017, 44(5): 443~450
www.pibb.ac.cn

基于有序条件互信息和有限父结点构建基因调控网络 *

刘 飞[1,2] 张绍武[1)**] 高红艳[2)]
([1)] 西北工业大学自动化学院，信息融合教育部重点实验室， 西安 710072；[2)] 宝鸡文理学院物理与光电技术学院，宝鸡 721016)

摘要 基因调控网络重建是功能基因组研究的基础，有助于理解基因间的调控机理，探索复杂的生命系统及其本质．针对传统贝叶斯方法计算复杂度高、仅能构建小规模基因调控网络，而信息论方法假阳性边较多、且不能推测基因因果定向问题．本文基于有序条件互信息和有限父结点，提出一种快速构建基因调控网络的 OCMIPN 算法．OCMIPN 方法首先采用有序条件互信息构建基因调控相关网络；然后根据基因调控网络拓扑先验知识，限制每个基因结点的父结点数量，利用贝叶斯方法推断出基因调控网络结构，有效降低算法的时间计算复杂度．人工合成网络及真实生物分子网络上仿真实验结果表明：OCMIPN 方法不仅能构建出高精度的基因调控网络，且时间计算复杂度较低，其性能优于 LASSO、ARACNE、ScanBMA 和 LBN 等现有流行算法．

关键词 基因调控网络，贝叶斯网络模型，有序条件互信息，有限父结点，因果定向
学科分类号 TP391 **DOI**: 10.16476/j.pibb.2016.0367

基因调控网络(gene regulatory networks, GRNs)是一个基因组内基因相互作用而形成的关系网络，它可以从基因作用角度揭示生命现象及其本质，是功能基因组学研究的重要内容，而基因调控网络的构建有助于理解基因间的调控机理、预测未知基因功能、认识疾病发病机理、加速药物研发[1–3]．基因芯片技术和高通量测序技术生产出大规模基因表达数据，从这些基因表达数据出发，目前已发展了许多计算方法[4–13]推断基因调控网络，这些方法可归类为：监督学习方法[8–9]、信息论方法[12–13]和模型方法[10–11]．监督学习方法可以高精度构建基因调控网络，但需要先验调控信息指导，而已知的基因调控标签较少．信息论方法可构建出大规模的基因网络，不需要先验调控信息，但构建的基因网络是一种相关网络，而非真正的基因调控网络，且假设不同时间点的样本序列独立．模型方法能够对基因调控动态行为提供更深层次理解，但模型参数对基因网络构建精度有较大影响，当构建的基因调控网络包含大量的基因时，需要花费大量时间来学习条件依赖关系，算法的时间复杂度较高．

模型方法一般包括常微分方程模型[10, 14]、多元线性回归模型[15–16]、线性规划模型[17–18]、布尔网络模型[19–20]和贝叶斯网络模型[6–7, 11]等．微分方程模型是一种简单的表达形式，所需模型参数较少，识别过程相对容易，但是这种模型构造的基因调控网络准确性较低．布尔网络模型较为简单，然而由于它是一种离散的数学模型，不能很好地反映细胞的实际情况．线性回归和线性规划模型是一种简单的线性数学模型，只能处理生物基因表达数据的线性关系，不能处理非线性关系，应用范围较小．另外，这些模型对数据噪声处理能力相对较弱，在基因表达数据缺失较多的情况下，这些模型无法高效构建调控基因网络．贝叶斯网络模型能够捕捉基因表达数据中固有的噪声，基于统计假设揭示基因表达水平中的因果关系，构建出较高精度的无环基因调控网络．但贝叶斯网络模型需要花费大量时间来学习

* 国家自然科学基金资助项目(91430111, 61473232, 61170134).
** 通讯联系人.
Tel: 029-88431308, E-mail: zhangsw@nwpu.edu.cn
收稿日期：2016-11-23，接受日期：2017-04-25

条件依赖关系，导致算法计算复杂度较高，不能用于大规模的基因调控网络构建．针对贝叶斯网络模型的高计算复杂度问题，人们一般采用贪婪爬山(greedy-hill climbing)、马尔科夫链蒙特卡洛(Markov Chain Monte Carlo)和模拟退火(simulated annealing)等启发式搜索算法减小计算复杂度，学习贝叶斯网络结构[21-22]．最近，研究者又提出了另外一些策略限制贝叶斯搜索范围，如ScanBMA[6]和LBN[7]算法根据先验知识选择候选调控基因，降低贝叶斯模型计算复杂度，但它们的候选调控基因选择范围仍然较大，其算法复杂度相对较高．

通过对已有基因调控网络拓扑结构分析发现[23-24]：基因结点的入度值呈指数下降，即基因调控网络中大部分基因结点被少数几个父基因结点调控，仅有极个别的基因结点被多个父结点基因调控．也就是说基因调控网络中某一基因的调控因子数目是有限的，这些有限的调控基因我们称之为“有限父结点”．鉴于此，本文提出一种基于有序条件互信息和有限父结点的网络构建算法(OCMIPN)，快速构建基因调控网络．OCMIPN算法从基因表达数据出发，首先应用互信息构建初始基因相关网络，然后基于有序条件互信息删除网络中的冗余基因关联边，最后采用限制父结点数量的策略学习贝叶斯网络结构，快速构建基因调控网络．OCMIPN算法可高精度构建基因调控网络，有效降低算法的计算复杂度．

1 数据集与方法

1.1 数据集

为了评价算法性能，本文在4个计算机模拟网络、1个人工合成网络和1个真实生物分子网络数据上验证OCMIPN算法的基因调控网络构建性能．4个计算机模拟网络(Net10、Net20、Net50、Net100)数据来自于DREAM竞赛数据[25]，该竞赛数据包含基因表达数据和标准网络，其网络是经实验验证了的酵母(Yeast)和大肠杆菌(*Escherichia coli*)调控网络．Net10，Net20，Net50和Net100网络分别包含10、20、50、100个基因和10、45、77、166条基因调控边．人工合成网络数据IRMA来自于文献[15]，该网络为酿酒酵母(*Saccharomyces cerevisiae*)合成网络，包含5个基因、6条基因调控边．真实生物分子网络数据为大肠杆菌SOS DNA修复网络数据[26]，包含9个基因，24条基因调控边．

1.2 互信息和条件互信息

互信息(mutual information，MI)不仅可以度量基因间的非线性相关性，且能够有效处理高维低样本基因表达数据，该度量和条件互信息(conditional mutual information，CMI)已广泛应用于基因网络构建．

若基因表达数据用向量 X(或 Y)表示，向量元素表示基因在不同时间或不同条件下的表达值，则基因变量 X 和 Y 之间相关性可用互信息 $MI(X, Y)$ 度量．

$$MI(X, Y)=-\sum_{x\in X,\, y\in Y} p(x, y)\log\frac{p(x, y)}{p(x),\, p(y)} \tag{1}$$

其中 $p(x)$表示基因变量 X 为 x 时的概率值，$p(x, y)$表示基因变量 X 和 Y 分别为 x 和 y 时的联合概率值．为方便计算，根据高斯核概率密度函数[27]，上式可以写为：

$$MI(X, Y)=\frac{1}{2}\log\frac{|C(X)|\cdot|C(Y)|}{|C(X, Y)|} \tag{2}$$

其中 C 表示基因变量的协方差矩阵，$|C|$ 表示矩阵 C 的行列式．如果基因变量 X 和 Y 相互独立，则互信息值 $I(X, Y)$为零．

条件互信息表示2个基因变量在第3个基因变量条件下的条件依赖性，基因变量 X 和 Y 在基因变量 Z 条件下的条件互信息 $CMI(X,Y|Z)$定义如下：

$$CMI(X,Y|Z)=\sum_{x\in X,\, y\in Y,\, z\in Z} p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{3}$$

其中 $p(x|z)$和 $p(y|z)$分别表示基因变量 X、Y 在基因Z条件下的概率，$p(x,y|z)$表示基因变量 X 和 Y 在基因 Z 条件下的联合概率，表示基因变量 X, Y 和 Z 的联合概率．为方便计算，类似公式(2)，公式(3)可简化为：

$$CMI(X,Y|Z)=\frac{1}{2}\log\frac{|C(X,Z)|\cdot|C(Y,Z)|}{|C(Z)|\cdot|C(X,Y,Z)|} \tag{4}$$

如果基因变量 X 和 Y 在变量 Z 条件下相互独立，则条件互信息值 $CMI(X,Y|Z)$为零．Z 为单个基因，则为一阶条件互信息；Z 为2个基因，则为二阶条件互信息，以此类推．考虑到算法的复杂度及本文所用数据特性，本文采用三阶条件互信息．

1.3 贝叶斯网络模型

对于一个随机变量 $X=\{X_1, X_2, \cdots, X_n\}$，贝叶斯网络(Bayesian Network，BN)是这些变量之间概率统计关系的一个图模型 G，它是一个有向无环图(directed acyclic graph，DAG)．贝叶斯网络中，结

点表示随机变量(基因)，边表示随机变量间的概率依赖．假设从结点 A 到结点 B 存在一条有向边，那么我们就称结点 A 是结点 B 的父亲，结点 B 是结点 A 的孩子．一个结点在给定父结点的情况下，根据 Markov 假设，这个结点和它的非子孙结点相互独立，可用 $P(X_1, X_1, \cdots, X_n)$联合概率分布来表示．基于图模型，$P(X_1, X_1, \cdots, X_n)$可分解为一系列条件概率的乘积：

$$P(X_1, X_2, \cdots, X_n)=\prod_{X_i \in X} P(X_i|Pa_i) \quad (5)$$

其中 Pa_i 为图 G 中结点 X_i 的父结点集合．

贝叶斯网络结构学习的目标就是基于训练数据 D，找到与数据 D 匹配程度最高的 BNs 结构．目前 BNs 结构学习方法可归类为基于约束的方法和基于打分搜索的方法．基于约束的方法就是通过条件独立性测试找出数据中隐含的条件独立关系，寻找与这些条件独立关系一致的网络结构，该类方法比较直观，但条件测试过多，且高阶测试误差较大[28]．基于打分搜索的方法通过打分函数，在搜索空间找出一个得分最高的结构，该类方法是一种统计驱动方法，搜索空间大，时间复杂度较高．其打分函数一般包括： 贝叶斯统计方法[29–30]、等价贝叶斯信息准则(BIC)方法[31]、最小描述长度(MDL)方法[32–33]和熵信息方法[34]．由于互信息测试(mutual information test，MIT)打分函数的良好性能，本文采用 MIT 打分函数(即贝叶斯联合概率)对 BNs 进行打分．

设 $X=\{X_1, X_2, \cdots, X_n\}$对应的样本分别为$\{r_1, r_2, \cdots, r_n\}$，数据集 D 中共有 N 个样本值，G 表示贝叶斯网络，$Pa_i=\{X_{i1}, X_{i2}, \cdots, X_{is_i}\}$表示结点 X_i 所有父结点集合，其对应的样本为$\{r_{i1}, r_{i2}, \cdots, r_{is_i}\}$，$s_i$ 为父结点个数，互信息测试打分函数定义如下[35]：

$$S_{MIT}(G:D)=\sum_{i=1,Pa_i\neq\varphi}^{n}\{2N\cdot MI(X_i,Pa_i)-\max_{\sigma_i}\sum_{j=1}^{s_i}\chi_{\alpha,l_i\sigma_i(j)}\} \quad (6)$$

$$l_i\sigma_i(j)=\begin{cases}(r_i-1)(r_{i\sigma_i(j)}-1)\prod_{k=1}^{j-1}r_{i\sigma_i(k)} & j=2,\cdots,s_i\\ (r_i-1)(r_{i\sigma_i(j)}-1) & j=1\end{cases} \quad (7)$$

$MI(X_i,Pa_i)$表示结点 X_i 和其父结点的互信息值，$\chi_{\alpha,l_i\sigma_i(j)}$表示显著性水平 α 下的卡方分布值，$\sigma_i=\{\sigma_i(1), \sigma_i(2), \cdots, \sigma_i(s_i)\}$为父结点 $Pa_i=\{X_{i1}, X_{i2}, \cdots, X_{is_i}\}$索引集合$\{1, 2, \cdots, s_i\}$的一个随机置换．

1.4 OCMIPN 算法

OCMIPN 算法采用有序条件互信息构建基因相关网络，然后通过限制每个基因结点的父结点数量构建基因调控网络，有效降低算法时间复杂度．OCMIPN 算法由三部分组成：a．基于互信息(MI)构建初始基因相关网络；b．基于有序条件互信息(CMI)删除冗余相关边；c．限制每个基因的父结点数量，基于贝叶斯网络模型构建基因调控网络，其流程如图 1 所示．

a．基于 MI 构建初始网络．OCMIPN 算法从基因表达数据出发，采用公式(2)计算所有基因对间的相关性，根据定义的 P-value 阈值，在大于此阈值的基因对上连边，形成初始基因相关网络 G_f，该网络包含一定数量的假阳性边．

b．基于有序 CMI 删除 G_f 网络冗余边．大量生物学实验表明：基因调控网络具有稀疏性，而互信息过高估计了基因间相关性，使得构建的基因相关网络包含较多的假阳性边，且互信息仅能定量描述基因间的两两相关关系，不能处理复杂的调控模式．条件互信息能处理 2 个或者 2 个以上的组合调控关系，因而可以采用 CMI 处理初始基因相关网络 G_f，以期发现 2 个以上基因间的调控关系，删除假阳性调控边．若直接采用公式(3)计算基因间的条件互信息，由于条件的顺序多样性，会产生不同 CMI 值，导致算法鲁棒性较差．因而，OCMIPN 算法采用有序 CMI 策略删除 G_f 网络冗余边，形成修正基因相关网络 G_p．有序 CMI 定义为：根据基因结点度大小，按降序对基因进行排序，将排序结果作为 CMI 的条件集合，从该集合中依次选取基因作为条件基因，计算 CMI．

c．基于有限父结点，采用贝叶斯网络模型构建 GRNs．图 1b 构建的基因网络 G_p 是一种无向网络，而真实基因调控网络是一种调控关系网络，一般可采用贝叶斯网络模型在此无向网络 G_p 上，计算所有基因间的条件依赖关系，构建 GRNs，但对于大规模的基因网络，贝叶斯网络模型需要花费大量时间学习基因间的条件依赖关系，算法时间复杂度较高．现有基因调控网络拓扑分析表明[23–24]：基因调控网络中基因结点的入度值呈指数级下降，即基因调控网络中大部分基因仅被少数几个基因(一般 1～3 个基因)调控，只有极个别的基因被多数基因调控．因而，OCMIPN 算法通过 maxPiter 策略[21]迭代限制每个基因的父结点数量，采用贝叶斯网络模型学习每个基因与其父结点基因间的条件依赖关系，构建基因调控网络．

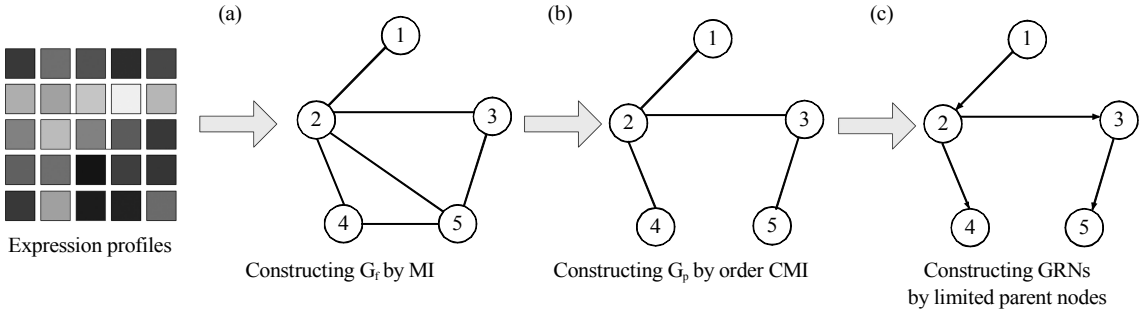

Fig. 1　Schematic diagram of OCMIPN method

1.5　系统评价指标

本文采用真阳性率(true positive rate，*TPR*)，假阳性率(flase positive rate，*FPR*)，阳性预测率(positive predictive，*PPV*)，错误发现率(flase discovery rate，*FDR*)，*F* 值，精确度(accuracy，*ACC*)和 matthews 相关系数(*MCC*)指标评价 OCMIPN 算法的网络构建性能，这些指标定义如下[7]：

$$TPR=R=TP/(TP+FN) \tag{8}$$

$$FPR=FP/(FP+TN) \tag{9}$$

$$FDR=FP/(FP+TP) \tag{10}$$

$$PPV=P=TP/(TP+FP) \tag{11}$$

$$ACC=(TP+TN)/(TP+FP+TN+FN) \tag{12}$$

$$F=2PPV\times TPR/(PPV+TPR) \tag{13}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{14}$$

其中，*TP* 为调控边的正确预测边数，*TN* 为非调控边的正确预测边数，*FP* 为非调控边误预测为调控边的数目，*FN* 为调控边误预测为非调控边的数目.

2　结果与讨论

为了检测 OCMIPN 算法在基因调控网络构建方面的性能，我们首先在小规模基因网络数据集(Net10、Net20、IRMA、SOS)上进行仿真验证，然后在较大规模尺度网络(Net50、Net100)上进行仿真验证，讨论实验结果、分析 OCMIPN 算法有效性及时间复杂度.

2.1　实验结果

首先在 Net10、Net20、IRMA 和 SOS 小规模基因网络数据集上，验证 OCMIPN 算法的基因调控网络构建性能，并与目前比较流行的算法 globalMIT[36]、LASSO[37]、ARACNE[13]、ScanBMA[6]和 LBN[7]比较．globalMIT[36]通过分解策略及互信息测试(mutual information test，MIT)打分度量，学习全局最优动态贝叶斯网络结构构建基因调控网络；LASSO[37]基于回归模型构建基因调控网络；ARACNE[13]基于信息论构建基因调控网络，而 ScanBMA[6]和 LBN[7]采用分解策略，搜寻有限基因父结点集合学习贝叶斯网络结构，快速构建基因调控网络．OCMIPN 方法和其他 5 种基因调控网络构建算法在 Net10、Net20、IRMA 和 SOS 数据集上的实验结果如表 1 所示．从表 1 中可以看出 OCMIPN 和 globalMIT 算法在 4 个小规模基因网络数据集上的预测精度、*MCC* 系数和 *F* 值远高于 LASSO、ARACNE、ScanBMA 算法，略高于 LBN 算法，但 OCMIPN 计算时间低于 LASSO、ARACNE、ScanBMA 和 LBN 算法，远低于 globalMIT 算法，例如 Net20 数据集上，OCMIPN 运行时间为 46.594 s，而 global 算法却需要 312.250 s．这些结果说明：globalMIT 算法可以高精度地构建小规模基因调控网络，因为 globalMIT 算法学习了全局最优贝叶斯网络结构；而 OCMIPN 算法不仅可以构建高精度的基因调控网络，而且运算时间较少，能够应用于构建较大规模的基因调控网络.

为了进一步说明 OCMIPN 算法可快速构建基因调控网络，我们在较大规模尺度网络 Net50 和 Net100 上验证 OCMIPN 算法性能．由于 globalMIT 算法的计算复杂度较高，比较适合构建 20 个结点以下的基因调控网络[21]，此处我们仅与 LASSO[37]、ARACNE[13]、ScanBMA[6]和 LBN[7] 4 种算法进行了比较．OCMIPN 算法及其他 4 种算法在 Net50 和 Net100 数据集上的实验结果见表 2．从表 2 可以看出，OCMIPN 和 LBN 算法在 Net50 和 Net100 网络

上的预测精度大于 LASSO、ARACNE 和 ScanBMA 算法，假阳性率较低，说明 OCMIPN 和 LBN 算法可以高精度地构建基因网络；OCMIPN 算法在 Net100 数据集的预测精度、*F* 值和 *MCC* 指标高于 LBN 算法，但在 Net 50 数据集上却低于 LBN 算法，且 OCMIPN 算法在 Net50 和 Net100 数据集上的运行时间明显小于 LBN 算法，说明相对于 LBN 算法，OCMIPN 算法更适合高精度地构建大规模基因调控网络；与 ScanBMA[6]快速贝叶斯网络构建算法相比，OCMIPN 算法不仅运行时间小于 ScanBMA，且基因网络构建的精度、*F* 值和 *MCC* 系数明显高于 ScanBMA，如 Net100 数据集上，OCMIPN 算法的 *F* 值、*MCC* 系数分别比 ScanBMA 算法高 0.089、0.088，说明 OCMIPN 算法的确可以高精度、快速地构建大规模基因调控网络.

Table 1 Results of six methods on the four small scale datasets of Net10、Net20、IRMA and SOS

Datasets	Methods	*ACC*	*MCC*	*F*	Runtime/s
Net10	globalMIT	0.940	0.730	0.750	17.482
	LASSO	0.880	–0.191	0.145	13.214
	ARACNE	0.889	0.619	0.643	11.052
	ScanBMA	0.930	0.629	0.667	8.945
	LBN	0.944	0.760	0.783	10.462
	OCMIPN	0.950	0.763	0.783	7.049
Net20	globalMIT	0.950	0.695	0.722	312.250
	LASSO	0.873	0.351	0.414	159.430
	ARACNE	0.910	0.542	0.581	120.436
	ScanBMA	0.920	0.536	0.579	85.401
	LBN	0.930	0.583	0.622	69.357
	OCMIPN	0.948	0.676	0.704	46.594
IRMA	globalMIT	0.800	0.369	0.444	0.232
	LASSO	0.600	0.016	0.286	0.232
	ARACNE	0.640	0.067	0.308	0.168
	ScanBMA	0.760	0.266	0.400	0.081
	LBN	0.800	0.369	0.444	0.033
	OCMIPN	0.800	0.369	0.444	0.012
SOS	globalMIT	0.778	0.495	0.653	15.470
	LASSO	0.469	0.074	0.442	14.198
	ARACNE	0.486	0.083	0.479	12.056
	ScanBMA	0.704	0.361	0.571	11.824
	LBN	0.736	0.413	0.612	9.482
	OCMIPN	0.778	0.467	0.625	8.9523

Table 2 Results of five methods on two large scale datasets of Net50 and Net100

Datasets	Methods	*TPR*	*FPR*	*FDR*	*PPV*	*F*	*ACC*	*MCC*	Runtime/s
Net50	LASSO	0.351	0.062	0.846	0.154	0.214	0.919	0.195	45.528
	ARACNE	0.584	0.040	0.676	0.324	0.417	0.949	0.411	81.426
	ScanBMA	0.364	0.051	0.811	0.183	0.249	0.931	0.229	42.527
	LBN	0.403	0.011	0.456	0.544	0.463	0.971	0.453	50.892
	OCMIPN	0.390	0.016	0.559	0.441	0.414	0.965	0.397	39.841
Net100	LASSO	0.181	0.051	0.943	0.057	0.087	0.936	0.074	154.621
	ARACNE	0.506	0.033	0.793	0.207	0.293	0.959	0.306	218.197
	ScanBMA	0.265	0.007	0.593	0.407	0.321	0.981	0.320	133.124
	LBN	0.283	0.005	0.510	0.489	0.359	0.983	0.364	157.942
	OCMIPN	0.350	0.006	0.504	0.496	0.410	0.983	0.408	126.453

2.2 有序条件互信息有效性分析

条件互信息(CMI)能处理 2 个或者 2 个以上的组合调控关系，可有效度量基因间的作用关系．但由于其条件顺序的多样性，会产生不同的 CMI 值，导致算法鲁棒性较差．因而，本文定义有序条件互信息，即根据基因结点度大小，按降序对基因进行排序，将排序结果作为 CMI 的条件集合，从该集合中依次选取基因作为条件基因，根据条件基因的排序，依次计算 CMI．为验证有序条件互信息的有效性，我们在 Net10 数据集上，分别通过随机选择基因条件顺序计算 CMI(随机 CMI)、以基因结点度大小顺序为基因条件顺序计算 CMI(有序 CMI)，采用 OCMIPN 构建基因调控网络，实验对比结果见表 3．从表 3 可以看出随机 CMI 的实验结果差异性较大，有序 CMI 的阳性预测值(*PPV*)、精确度(*ACC*)、*F* 值和 *MCC* 系数大于随机 CMI，且错误率(*FDR*)最低，说明有序条件互信息策略可有效删除假阳性边，提高基因网络构建精度．

Table 3 Results of OCMIPN with order CMI and random CMI strategies on the Net10 dataset

Methods	*FDR*	*PPV*	*F*	*ACC*	*MCC*
Random CMI_1	0.25	0.75	0.82	0.91	0.77
Random CMI_2	0.36	0.64	0.75	0.87	0.68
Random CMI_ 3	0.38	0.62	0.70	0.62	0.60
Random CMI_4	0.13	0.88	0.78	0.91	0.73
Order CMI	0.10	0.90	0.90	0.95	0.87

2.3 OCMIPN 算法时间复杂度分析

OCMIPN 算法时间复杂度包括 3 部分：第一部分，基于 MI 构建基因相关网络 G_f，算法最大时间复杂度为 $O(n^2)$，其中 n 为网络所包含的基因数目；第二部分，基于有序 CMI 删除 G_f 网络冗余边，需要对 n 个结点排序，然后采用 CMI 去除网络中的冗余边，算法最大时间复杂度也为 $O(n^2)$；第三部分，基于有限父结点利用贝叶斯网络模型构建 GRNs，此过程要为每个基因结点选择合适的调控父结点，其最大时间复杂度为 $O(n\times 2^m)$，m 为基因父结点数量．因此，OCMIPN 算法总时间复杂度为 $O(n^2+n^2+n\times 2^m)$，由于 m 远远小于 n，则 OCMIPN 算法的时间复杂度为 $O(n^2)$，远低于传统贝叶斯网络构建方法的时间复杂度 $O(2^n)$．

3 结　论

利用贝叶斯网络推断基因调控网络的最优结构是一个 NP-hard 问题，虽然目前基于约束和打分搜索策略学习贝叶斯网络结构可以在一定程度上降低搜索空间，但对于较大规模的基因调控网络，算法时间复杂度仍然较高．基因调控网络分析表明：基因结点的入度值呈指数下降，即基因调控网络中大部分基因仅被少数几个基因(一般 1～3 个基因)调控，仅有极个别的基因结点被多个父结点基因调控．因此，为了降低基因调控网络构建算法的计算时间复杂度，人们采用限制每个基因结点的父结点数量策略构建基因调控网络．另外，采用 MI 和 CMI 构建基因相关网络初步确定基因间的相关关系，可有效降低贝叶斯网络结构学习的搜索空间．

针对传统贝叶斯方法计算复杂度高，仅能构建小规模基因调控网络，而现有快速贝叶斯模型构建大规模基因调控网络精度低、时间复杂度相对较高问题，本文基于有序条件互信息和有限父结点，提出一种快速构建基因调控网络的 OCMIPN 算法．OCMIPN 算法从基因表达数据出发，首先利用互信息构建初始基因相关网络，然后基于有序条件互信息删除网络中的冗余基因关联边，最后采用限制父结点数量策略学习贝叶斯网络结构，构建基因调控网络．与现有一些性能较优的基因调控网络算法相比，OCMIPN 算法可以高精度、快速构建大规模基因调控网络；有序条件互信息可有效删除假阳性边、提高基因网络构建精度；限制父结点数量能够有效降低算法时间复杂度．虽然 OCMIPN 算法可以快速、有效地构建较大规模的基因调控网络，但如何确定父结点基因数量上限、如何进一步降低算法时间复杂度使其能够构建更大规模的基因调控网络，仍需要进一步深入研究．

参 考 文 献

[1] Repsilber D, Liljenström H, Andersson S G E. Reverse engineering

of regulatory networks: simulation studies on a genetic algorithm approach for ranking hypotheses. Biosystems, 2002, **66**(1): 31−41

[2] Madhamshettiwar P B, Maetschke S R, Davis M J, *et al*. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Medicine, 2012, **4**(5): 1−16

[3] Zhao J, Zhou Y, Zhang X, *et al*. Part mutual information for quantifying direct associations in networks Proc Natl Acad Sci USA, 2016, **113**(18): 5130−5135

[4] 徐肖江, 王连水, 丁达夫. 从酵母表达时间序列估计基因调控网络. 生物化学与生物物理学报, 2003, **35**(8): 707−716
Xu X J, Wang L S, Ding D F. Prog Biochem Biophys, 2003, 35(8): 707−716

[5] Yalamanchili H K, Yan B, Li M J, *et al*. DDGni: dynamic delay gene-network inference from high-temporal data using gapped local alignment. Bioinformatics, 2014, **30**(3): 377−383

[6] Young W C, Raftery A E, Yeung K Y. Fast Bayesian inference for gene regulatory networks using ScanBMA. Bmc Systems Biology, 2014, **8**(1): 47−47

[7] Liu F, Zhang S W, Guo W F, *et al*. Inference of gene regulatory network based on local bayesian networks. PLoS Comput Biol, 2016, **12**(8): e1005024

[8] Kotera M, Yamanishi Y, Moriya Y, *et al*. GENIES: gene network inference engine based on supervised analysis. Nucleic Acids Research, 2012, **40**(W1): W162−W167

[9] Mordelet F, Vert J P. SIRENE: supervised inference of regulatory networks. Bioinformatics, 2008, **24**(16): i76−i82

[10] Tian T, Burrage K. Stochastic models for regulatory networks of the genetic toggle switch. Proc Natl Acad Sci USA, 2006, **103**(22): 8372−8377

[11] Zou M, Conzen S D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics, 2005, **21**(1): 71−79

[12] Zhang X, Zhao J, Hao J K, *et al*. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic Acids Research, 2015, **43**(5): e31−e31

[13] Margolin A A, Nemenman I, Basso K, *et al*. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 2006, **7**(1): S7

[14] Sakamoto E, Iba H. Inferring a system of differential equations for a gene regulatory network by using genetic programming. Proceedings of the 2001 Congress on Evolutionary. IEEE Press, 2001: 720−726

[15] Cantone I, Marucci L, Iorio F, *et al*. A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. Cell, 2009, **137**(1): 172−181

[16] Honkela A, Girardot C, Gustafson E H, *et al*. Model-based method for transcription factor target identification with limited data. Proc Natl Acad Sci USA, 2010, **107**(17): 7793−7798

[17] Zhang X, Liu K, Liu Z P, *et al*. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics, 2013, **29**(1): 106−113

[18] Zhu H, Rao R S P, Zeng T, *et al*. Reconstructing dynamic gene regulatory networks from sample-based transcriptional data. Nucleic Acids Research, 2012, **40**(21): 10657−10667

[19] Shmulevich I, Dougherty E R, Kim S, *et al*. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics, 2002, **18**(2): 261−274

[20] Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pacific Symposium on Biocomputing, 1999, **4**: 17−28

[21] Nair A, Chetty M, Wangikar P P. Improving gene regulatory network inference using network topology information. Molecular BioSystems, 2015, **11**(9): 2449−2463

[22] Wu J, Zhao X, Lin Z, *et al*. Large scale gene regulatory network inference with a multi-level strategy. Molecular BioSystems, 2016, **12**(2): 588−597

[23] Albert R. Scale-free networks in cell biology. Journal of Cell Science, 2005, **118**(21): 4947−4957.

[24] Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization. Nature Reviews Genetics, 2004, **5**(2): 101−113

[25] Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics, 2011, **27**(16): 2263−2270

[26] Shen-Orr S S, Milo R, Mangan S, *et al*. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature Genetics, 2002, **31**(1): 64−68

[27] Basso K, Margolin A A, Stolovitzky G, *et al*. Reverse engineering of regulatory networks in human B cells. Nature Genetics, 2005, **37**(4): 382−390

[28] 贾海洋, 刘大有, 陈 娟,等. 免疫遗传算法学习贝叶斯网等价类. 吉林大学学报(理学版), 2009, **47**(1): 48−56
Jia H Y, Liu D Y, Chen J, *et al*. Journal of Jilin University(Science Edition), 2009, **47**(1): 48−56

[29] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 1992, **9**(4): 309−347

[30] Heckerman D, Geiger D, Chickering D M. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 1995, **20**(3): 197−243

[31] Schwarz G. Estimating the dimension of a model. The Annals of Statistics, 1978, **6**(2): 461−464

[32] Hansen M H, Yu B. Model selection and the principle of minimum description length. Journal of the American Statistical Association, 2001, **96**(454): 746−774

[33] Lam W, Bacchus F. Learning Bayesian belief networks: An approach based on the MDL principle. Computational Intelligence, 1994, **10**(3): 269−293

[34] Chow C, Liu C. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 1968, **14**(3): 462−467

[35] Campos L M. A scoring function for learning bayesian networks

based on mutual information and conditional independence tests. Journal of Machine Learning Research, 2006, **7**(2): 2149-2187

[36] Vinh N X, Chetty M, Coppel R, *et al*. GlobalMIT: learning globally optimal dynamic bayesian network with the mutual information test criterion. Bioinformatics, 2011, **27**(19): 2765-2766

[37] Geeven G, van Kesteren R E, Smit A B, *et al*. Identification of context-specific gene regulatory networks with GEMULA——gene expression modeling using LAsso. Bioinformatics, 2012, **28** (2): 214-221

Inferring Gene Regulatory Networks Based on Ordered Conditional Mutual Information and Limited Parent Nodes*

LIU Fei[1,2], ZHANG Shao-Wu[1)**], GAO Hong-Yan[2)]

([1)] *Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an* 710072, China;

[2)] *Institute of Physics and Optoelectronics Technology, Baoji University of Arts and Science, Baoji* 721016, China)

Abstract Inferring the gene regulatory networks (GRNs) structure is the research basis of functional genomics. GRNs can help to understand the regulatory mechanism among genes, exploring the essence of complex life system. Traditional Bayesian network methods cannot handle large-scale networks due to their high computational complexity, while information theory-based methods cannot identify the directions of regulatory interactions and also suffer from false positive/negative problems. By using the ordered conditional mutual information (CMI) and limited parent node genes, in this work, we present a novel algorithm (namely OCMIPN) to fast infer GRNs from gene expression data. OCMIPN first uses ordered conditional mutual information to construct an initial GRN relation network. Then, according to the priori knowledge of gene regulatory network topology structure, BN method is employed to generate final GRNs by limiting the number of parent nodes for each gene, which significantly reduces the computational complexity. Tested on the synthetic networks as well as real biological molecular networks with different sizes and topologies, the results show that OCMIPN can infer RGNs with higher accuracy and low computational times. The OCMIPN's performance outperforms other state-of-the-art methods, such as LASSO, ARACNE, ScanBMA and LBN.

Key words gene regulatory network, Bayes network model, ordered conditional mutual information, limited parent nodes, causality orientation

DOI: 10.16476/j.pibb.2016.0367

*This work was supported by a grant from The National Natural Science Foundation of China (91430111, 61473232, 61170134).

**Corresponding author.

Tel: 86-29-88431308, E-mail: zhangsw@nwpu.edu.cn

Received: November 23, 2016 Accepted: April 25, 2017