基于部分互信息和贝叶斯打分函数的基因调控网络构建算法

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, 陕西 西安 710072
2.宝鸡文理学院 物理与光电技术学院, 陕西 宝鸡 721016)

摘 要:从基因表达数据出发重构基因调控网络,可有效挖掘基因间调控关系,深层次地理解生物调控过程。传统的相关性系数模型、偏相关系数模型仅能发现基因间线性关系,而互信息和条件互信息可用于发现基因间的非线性关系,且能够处理高维低样本基因表达数据。但互信息过高估计基因间的相关性,条件互信息过低估计基因间的相关性,从而导致推断出的基因网络假阳性率和假阴性率较高,且不能推断基因调控方向。因而,基于部分互信息和贝叶斯打分函数,提出一种新的基因调控网络构建算法(命名为 PMIBSF)。基于部分互信息,PMIBSF 算法首先删除初始基因相关网络中的冗余关联边,然后采用贝叶斯网络互信息测试打分函数学习贝叶斯网络结构,快速构建基因调控网络。在计算机模拟网络和真实生物分子网络上,仿真实验结果表明:PMIBSF 性能优于目前较流行的 LP、PC-alg、NARROMI 和 ARACNE 算法,可高精度构建基因调控网络。

关 键 词:部分互信息;互信息测试打分;贝叶斯网络;协方差矩阵;基因调控网络
中图分类号:TP391 **文献标志码**:A **文章编号**:1000-2758(2017)05-0876-08

随着基因芯片技术和高通量测序技术的发展,产生了大量的基因表达数据。从这些基因表达数据出发构建复杂的基因调控网络(gene regulatory networks,GRNs),可有效描述一个物种或者组织内基因间的相互作用关系,进而在生物分子网络角度认识生命现象并揭示生命活动的基本规律,GRNs 重构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]。近年来涌现出许多基因调控网络的重构方法模型,如皮尔逊相关系数模型[1]、常微分方程模型[4]、随机网络模型[5]、布尔网络模型[6]、回归模型[7]、线性规划模型[8]和贝叶斯网络模型[9]等,每一种模型方法都有各自的优点和局限性。

皮尔逊相关系数(pearson correlation coefficient,PCC)模型是一种最常用的基因调控网络构建方法,它可以发掘基因变量间的线性相关性,但是不能识别基因变量间是直接调控还是间接调控。偏相关性(partial correlation,PC)由于考虑了额外的相关信息可以克服 PCC 模型的局限性,发现基因变量间的直接调控关系。Barzel 等人[10]应用 PC 指标构建了一种动态相关性基因调控网络,消除网络中基因间的间接影响,从而区分基因间的直接调控和间接调控。但是这两种方法只能识别基因变量间的线性相关性,不能识别基因变量间的非线性相关性,而在真实生物分子网络中基因间的非线性相关性扮演了一个很重要的角色。互信息(mutual information,MI)和条件互信息(conditional mutual information,CMI)不仅可以从高维低样本基因表达数据中发现基因变量间的线性关系,也可以发现基因变量间的非线性关系。因而,基于互信息和条件互信息,涌现出了许多基因调控网络构建方法[11-12]。但是互信息过高地估计了基因变量之间的相互作用关系,导致推断出的基因网络有较多的假阳性边;而条件互信息往往过低地估计了基因变量之间的相关性,从而导致推

收稿日期:2017-03-01 **基金项目**:国家自然科学基金(91430111、61473232、61170134)资助
作者简介:刘飞(1981—),西北工业大学博士研究生,主要从事生物信息学研究。

断出的基因网络有较高的假阴性边。鉴于此,Zhao 等人[13] 提出采用部分互信息(part mutual information,PMI)度量基因变量间的相关性,但 PMI 不能确定基因间的调控关系,而基因间调控方向的确定可通过贝叶斯打分策略来实现。贝叶斯打分策略通过打分函数,在搜索空间找出一个得分最高的网络结构。打分函数一般包括:贝叶斯统计方法[14-15]、等价贝叶斯信息准则(BIC)方法[16]、最小描述长度(MDL)方法[17-18]、熵信息方法[17] 和互信息测试打分函数(MIT)等。MIT 是一种非常通用打分测试函数,利用显著性水平下的卡方分布值来估算基因结点间的调控关系,具有搜索空小、时间复杂度低等优点。

针对基于互信息和条件互信息构建的基因网络假阳性/假阴性率高、偏互信息不能确定基因调控方向问题,本文从基因表达数据出发,首先构建基因完全图,然后采用部分互信息构建稀疏基因相关网络,采用贝叶斯网络互信息测试打分函数确定基因调控方向,提出一种新的基因调控网络构建算法(命名为 PMIBSF),快速构建基因调控网络,并在计算机模拟网络和真实生物基因调控网络上验证 PMIBSF 算法性能。

1　理论方法

1.1　互信息和条件互信息

基于信息论的互信息和条件互信息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,广泛被用来构建基因相关网络,且效果较好。

若基因表达数据用向量 $\boldsymbol{X}$(或 $\boldsymbol{Y}$) 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 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间相关性可用互信息 $MI(\boldsymbol{X},\boldsymbol{Y})$ 度量。

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

式中,$p(x)$ 表示基因变量 $\boldsymbol{X}$ 为 x 时的概率值,$p(x,y)$ 表示基因变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 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[19],上式可以写为:

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

式中,$\boldsymbol{C}$ 表示基因变量的协方差矩阵,$|\boldsymbol{C}|$ 表示矩阵 $\boldsymbol{C}$ 的行列式。如果基因变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 相互独立,则互信息值 $MI(\boldsymbol{X},\boldsymbol{Y})$ 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 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在基因变量 $\boldsymbol{Z}$ 条件下的条件互信息 $CMI(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z})$ 定义如下:

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

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

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

如果基因变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在变量 $\boldsymbol{Z}$ 条件下相互独立,则条件互信息值 $CMI(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z})$ 为零。

1.2　部分互信息

互信息在测量 2 个变量之间的相关性时,往往过高地估计了两者之间的相关性,从而导致了网络有较高的假阳性边。条件互信息在测量 2 个变量之间的相关性时,往往过低地估计了两者之间的相关性,从而导致了网络有较高的假阴性边。为此,Zhao 等人[13] 提出采用部分互信息(part mutual information,PMI)[20] 降低假阳率和假阴性率。

假设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 表示 2 个随机变量,如果它俩彼此独立,它们之间的相关性为零,则

$$p(x)p(y) = p(x,y) \quad (5)$$

同理如果随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在变量 $\boldsymbol{Z}$ 条件下彼此独立,则

$$p(x|z)p(y|z) = p(x,y|z) \quad (6)$$

随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在给定变量 $\boldsymbol{Z}$ 条件下的部分独立性定义如下[20]:

$$p^*(x|z)p^*(y|z) = p(x,y|z) \quad (7)$$

式中,$p^*(x|z) = \sum_y p(x|z,y)p(y)$,$p^*(y|z) = \sum_x p(y|z,x)p(x)$。

根据部分独立性公式(7) 和 KL 距离(Kullback-Leibler divergence)[21],PMI 定义如下:

$$PMI(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}) = D(p(x,y,z)\|p^*(x|z)p^*(y|z)p(z)) \quad (8)$$

式中,$p(x,y,z)$ 表示基因变量 $\boldsymbol{X}$、$\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ 联合概率分

布，$D(p(x,y,z)\|p^*(x\mid z)p^*(y\mid z)p(z))$ 表示 $p(x,y,z)$ 到 $p^*(x\mid z)p^*(y\mid z)p(z)$ 的 KL 距离，PMI 的公式也可以写成[20]：

$$PMI(\boldsymbol{X},\boldsymbol{Y}\mid\boldsymbol{Z})=\sum_{x,y,z}p(x,y,z)\log\frac{p(x,y\mid z)}{p^*(x\mid z)p^*(y\mid z)} \tag{9}$$

根据公式(1)和(7)，PMI 的公式进一步可写为[20]：

$$PMI(\boldsymbol{X},\boldsymbol{Y}\mid\boldsymbol{Z})=CMI(\boldsymbol{X},\boldsymbol{Y}\mid\boldsymbol{Z})+D(p(x\mid z)\|p^*(x\mid z))+D(p(y\mid z)\|p^*(y\mid z)) \tag{10}$$

1.3 贝叶斯网络的互信息测试打分策略

对于一个随机变量 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_n\}$，贝叶斯网络(Bayesian network，BN) 就是这些变量之间概率统计关系的一个图模型，而且它是一个有向无环图 G。在贝叶斯网络中，顶点(结点) 就是随机变量(基因)，边就是随机变量(基因) 之间的概率依赖。假如在结点 A 到结点 B 有一条有向边，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，根据马尔可夫(Markov) 假设，这个结点和它的非子孙结点都是相互独立的，可以用 $\boldsymbol{P}(\boldsymbol{X}_1,\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)$ 这个联合概率分布来表示。基于图模型，它可以分解为一系列条件概率的乘积，表达式如下：

$$\boldsymbol{P}(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)=\prod_{X_i\in X}P(\boldsymbol{X}_i\mid Pa(\boldsymbol{X}_i)) \tag{11}$$

式中，$Pa(X_i)$ 为图 $\boldsymbol{G}$ 中结点 $\boldsymbol{X}_i$ 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 $\boldsymbol{D}$，找到与数据 $\boldsymbol{D}$ 匹配程度最高的贝叶斯网络结构。目前，贝叶斯网络结构学习可通过约束方法和打分搜索方法实现。鉴于互信息测试(mutual information test，MIT)[22] 打分函数的良好性能，常用来对贝叶斯网络结构进行打分。

设 $\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$ 对应的样本分别为 $\{r_1,r_2,\cdots,r_n\}$，数据集 $\boldsymbol{D}$ 中共有 $\boldsymbol{N}$ 个样本，$\boldsymbol{G}$ 表示贝叶斯网络，$\boldsymbol{Pa}_i=\{\boldsymbol{X}_{i1},\boldsymbol{X}_{i2},\cdots,\boldsymbol{X}_{is_i}\}$ 表示结点 $\boldsymbol{X}_i$ 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[22]：

$$S_{MIT}(\boldsymbol{G};\boldsymbol{D})=\sum_{i=1,Pa_i\neq\phi}^{n}\left\{2\boldsymbol{N}\cdot MI(\boldsymbol{X}_i,\boldsymbol{Pa}_i)-\max_{\sigma_i}\sum_{j=1}^{s_i}\chi_{\alpha,l_i\sigma_i(j)}\right\} \tag{12}$$

$$l_i\sigma_i(j)=\begin{cases}(r_i-1)(r_{i\sigma_i(j)}-1)\prod_{k=1}^{j-1}r_{i\sigma_i(k)} & j=2,\cdots,s_i\\(r_i-1)(r_{i\sigma_i(j)}-1) & j=1\end{cases} \tag{13}$$

$MI(\boldsymbol{X}_i,\boldsymbol{Pa}_i)$ 表示结点 $\boldsymbol{X}_i$ 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 $\boldsymbol{Pa}_i=\{\boldsymbol{X}_{i1},\boldsymbol{X}_{i2},\cdots,\boldsymbol{X}_{is_i}\}$ 索引集合 $\{1,2,\cdots,s_i\}$ 的一个随机置换。

1.4 PMIBSF 算法

PMIBSF 算法由四部分组成：1) 生成初始基因完全图；2) 运用部分互信息稀疏化基因完全图；3) 利用互信息测试打分函数对所有可能的基因网络结构打分；4) 确定得分最大的基因网络结构为最终的基因调控网络。图 1 为 PMIBSF 算法流程框图，其计算步骤如下：

1) 根据表达基因个数生成初始完全网络图 $\boldsymbol{G}_f$。

2) 运用部分互信息(PMI) 对初始基因完全图进行稀疏化。基因调控网络具有小世界性，是一种典型的稀疏网络。基因对之间的相关性计算指标可以计算所有基因对之间的相关性，如果基因对之间具有显著性较低的相关性值，则删除初始基因相关网络 $\boldsymbol{G}_f$ 中对应的网络边，以此对 $\boldsymbol{G}_f$ 网络进行稀疏化。避免了互信息、条件互信息的假阳性率和假阴性率高的问题，我们采用 PMI 度量基因间相关性，根据 P-value 的显著性，删除 $\boldsymbol{G}_f$ 网络中冗余的假阳性边，生成稀疏基因相关网络 $\boldsymbol{G}_p$。

3) 基于贝叶斯网络测试打分函数(MIT) 对稀疏基因相关网络 $\boldsymbol{G}_p$ 打分。基因相关网络 $\boldsymbol{G}_p$ 是一种无向网络，不能确定基因间的调控关系，而基因间调控方向的确定可通过贝叶斯 MIT 打分策略来实现，我们对网络 $\boldsymbol{G}_p$ 所有可能的拓扑结构用 MIT 进行打分排序。

4) 确定最终基因调控网络 $\boldsymbol{G}_l$。找出得分最大的网络图作为最终的基因调控网络 $\boldsymbol{G}_l$。

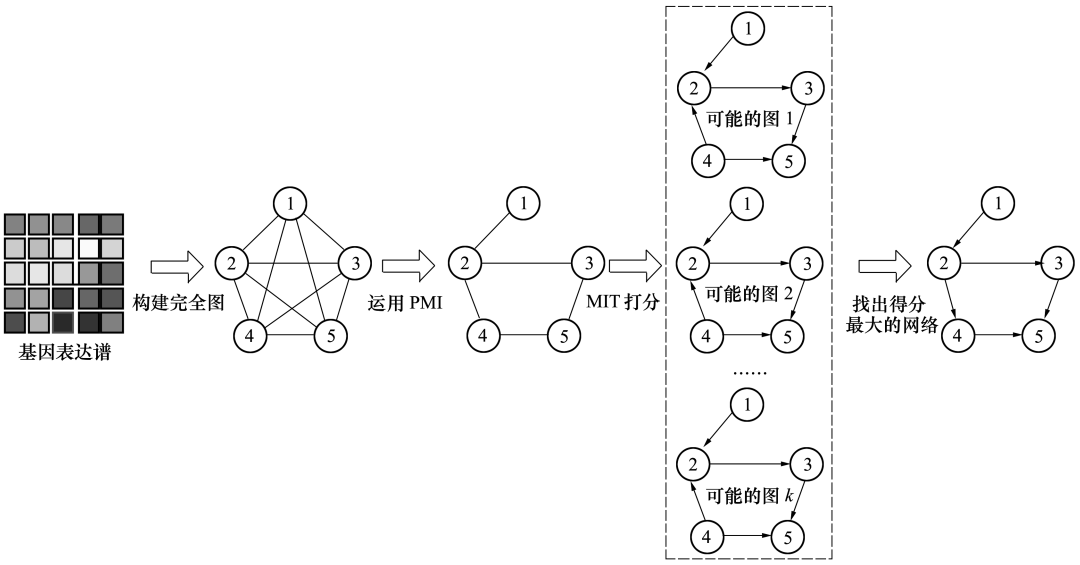

图 1　PMIBSF 算法流程图

2　结果与讨论

2.1　数据集和评价指标

为了评价算法性能,本文在 3 个计算机模拟网络、1 个人工合成网络和 1 个真实生物基因网络数据上,验证 PMIBSF 算法构建基因调控网络性能。计算机模拟网络(data10,data50,data100)数据来自于 DREAM 竞赛数据[23],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。data10、data50 和 data100 网络分别包含 10、50、100 个基因和 10、77、166 条基因调控边。人工合成网络数据 IRMA 来自于文献[24],该网络为酿酒酵母(yeast 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[25],包含 9 个基因,24 条基因调控边。

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 PMIBSF 算法性能,这些指标定义如下[3]:

$$TPR = R = TP/(TP+FN)$$

$$FPR = FP/(FP+TN)$$

$$FDR = FP/(FP+TP)$$

$$PPV = P = \frac{TP}{(TP+FP)}$$

$$ACC = (TP+TN)/(TP+FP+TN+FN)$$

$$F = 2PPV * TPR/(PPV+TPR)$$

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

式中,TP 表示调控边的正确预测数目,TN 表示非调控边的正确预测数目,FP 表示真实非调控边误预测为调控边的数目,FN 表示真实调控边误预测为非调控边的数目。

2.2　部分互信息性能分析

我们首先通过图 2 中的简单示例对比分析互信息(MI)、条件互信息(CMI)和部分互信息(PMI)构建基因相关网络的优劣性,然后在合成生物网络上进一步说明部分互信息(PMI)构建基因相关网络的优越性。

图 2 中,结点 $\boldsymbol{X}$、$\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ 表示 3 个基因变量,它们之间的线表示基因间的相互作用关系,线的粗细表示相关性的强弱。图 2a)中,基因变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 彼此是独立的(即它俩的相关性为零),它们和变量 $\boldsymbol{Z}$ 都有一定的相关性,运用互信息公式(2)计算结点 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相关性,其互信息值大于零,实际上 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 没有相互作用关系,即它间的相关性应该为零,而 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间 PMI 值为零;可见 MI 过高地估计了变量

之间的关系,而 PMI 能够正确地预测变量 X 和 Y 之间的相关性。图2b) 中,基因变量 X 和 Y 是相关的,变量 X 和 Z 之间的相关性较强,变量 Y 和 Z 之间的相关性较弱,通过计算可知 X 和 Y 之间的 CMI 值等于零,实际上 X 和 Y 的相关性不为零,而 X 和 Y 间的 PMI 值大于零;可见 CMI 过低地估计变量 X 和 Y 之间的关系,而 PMI 能够正确地预测变量 X 和 Y 之间的相关性。

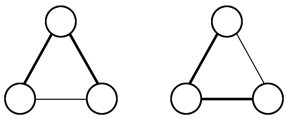

图 2 基因相关网络(线的粗细表示基因间相关性的强弱)

我们进一步采用 MI、CMI 和 PMI 3 种方法对 IRMA 数据构建基因调控网络,仿真验证结果见表 1。MI 方法和 PMI 方法的真阳率指标都高达 0.833,而 CMI 方法的真阳率指标却只有 0.667,这是因为 CMI 过低地估计了基因变量之间的作用关系,导致一些真实的调控边被遗漏。在假阳性率(FPR)方面,基于 MI 的方法取得了最差的效果,这是因为互信息过高地估计了基因变量之间的相互作用关系,导致推断出的基因网络有较多的假阳性边。在错误发现率(FDR)、阳性预测率(PPV)、F 值、精确度(ACC)和 Matthews 相关系数等方面的指标上,基于 PMI 的方法都取得了最好的结果,这充分证明了部分互信息是一种比较有效的基因间相关性度量指标。

表 1 MI、CMI 和 PMI 3 种指标在 IRMA 数据上的实验结果比对

指标	TPR	FPR	FDR	PPV	F	ACC	MCC
MI	0.833	0.571	0.615	0.385	0.526	0.550	0.252
CMI	0.667	0.500	0.636	0.364	0.471	0.550	0.154
PMI	0.833	0.357	0.500	0.500	0.625	0.700	0.436

2.3 实验结果与分析

首先在 deata10 基因网络模拟数据集上,验证 PMIBSF 算法的基因调控网络构建性能,并与 Zhao[13] 方法进行比较。图 3a) 是 data10 数据的标准网络,图 3b) 是 Zhao[13] 方法构建的基因相关网络(无向网络),预测出 9 条正确的相关边,没能预测出基因 4 和 9 之间的相关边。图 3c) 为我们 PMIBSF 算法构建的基因调控网络,PMIBSF 算法不仅预测出了 9 条基因相关边,且正确预测出 7 条有向调控边,说明 PMIBSF 算法可以较为准确地构建基因调控网络。

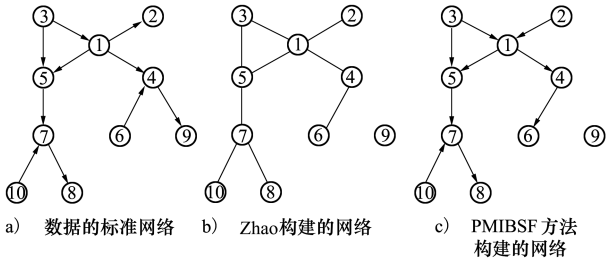

图 3 不同方法构建 data10 基因网络

然后,在 data50 和 data100 基因网络模拟数据集上,验证 PMIBSF 算法的基因调控网络构建性能,并与目前比较流行的算法 LP[26]、PC-alg[27]、NARROMI[11] 和 ARACNE[12] 比较。LP 是一种线性规划基因调控网络构建模型,利用目标函数优化问题使得构建的网络稀疏性和可靠性较高;PC-alg 是一种基于路径一致贪婪迭代的基因调控网络构建算法,其运行速度较快;而 NARROMI 和 ARACNE 都是基于信息论的基因调控网络构建方法。PMIBSF 方法和其他 4 种基因调控网络构建算法在 data50 和 data100 数据集上的实验结果如表 2 所示。从表中可以看出在 data50 数据集上,PMIBSF 和 ARACNE 算法的构建精度大于 LP、PC-alg 和 NARROMI 算法,且假阳性率较低,在错误发现率(FDR),F 值和 Matthews 相关系数(MCC)等指标方面也明显高于其他 3 种方法,这说明 PMIBSF 和 ARACNE 算法可以高精度地构建基因网络。在 data100 数据集上,PMIBSF 算法在各项性能指标上明显大于 LP、PC-alg、NARROMI 和 ARACNE 算法,取得了较好的性能,在 PPV、F 值、ACC 和 MCC 系数指标上,PMIBSF 算法比 ARACNE 算法分别高了 0.276、0.104、0.024 和 0.089。在 data50 数据集上 ARACNE 算法的 F 值指标比 PMIBSF 算法略高一些,错误发现率(FDR)却比 PMIBSF 算法的差一些,这是因为 ARACNE 方法采用的互信息过高地估计了基因变量之间的相关性所导致,但在 data100 数据集上 PMIBSF 算法的 F 值指标却明显高于 ARACNE 算法,这说明在中大规模基因调控网络构建上 PMIBSF 算法更为有效。

表 2　data50 和 data100 大规模基因网络数据集上 5 种算法实验结果比对

数据集	方法	TPR	FPR	FDR	PPV	F 值	ACC	MCC
	LP	0.389	0.085	0.870	0.130	0.195	0.899	0.182
	PC-alg	0.428	0.071	0.711	0.289	0.345	0.898	0.299
data50	NARROMI	0.532	0.062	0.783	0.217	0.308	0.925	0.307
	ARACNE	0.584	0.040	0.676	0.324	0.417	0.949	0.411
	PMIBSF	0.377	0.016	0.567	0.433	0.403	0.965	0.386
	LP	0.404	0.046	0.870	0.130	0.196	0.944	0.206
	PC-alg	0.457	0.026	0.624	0.376	0.142	0.956	0.392
data100	NARROMI	0.277	0.010	0.676	0.324	0.299	0.978	0.289
	ARACNE	0.506	0.033	0.793	0.207	0.293	0.959	0.306
	PMIBSF	0.337	0.006	0.517	0.483	0.397	0.983	0.395

最后,为了进一步说明 PMIBSF 算法构建基因调控网络的精确性,在真实生物分子 SOS 大肠杆菌网络数据上进行了验证,并与 LP、PC-alg、NARROMI 和 ARACNE 4 种算法进行了比较,详细的实验结果见表 3。在真阳率方面,PMIBSF 算法高于 LP 和 PC-alg 算法,却低于 NARROMI 和 ARACNE 算法,这是由于 NARROMI 和 ARACNE 算法都是基于 MI 的方法,推断出的基因网络冗余边数较多,使得 NARROMI 和 ARACNE 算法构建的网络真阳率指标较好,但假阳率也过高。在假阳率方面 PMIBSF 算法取得较好的结果,但是却高于 LP 算法,这是由于 LP 算法构建的 GRNs 过稀疏所导致。在其他的 PPV、F 值、ACC 和 MCC 系数等指标上,PMIBSF 算法都取得了最好的效果,这说明 PMIBSF 算法在真实生物分子网络上也可以高精度地构建基因网络。

表 3　SOS 基因网络数据集上 5 种算法实验结果比对

数据集	方法	TPR	FPR	FDR	PPV	F 值	ACC	MCC
	LP	0.208	0.146	0.583	0.417	0.278	0.639	0.079
	PC-alg	0.500	0.375	0.600	0.400	0.444	0.583	0.120
SOS	NARROMI	0.667	0.458	0.579	0.421	0.516	0.583	0.197
	ARACNE	0.708	0.625	0.638	0.362	0.479	0.486	0.083
	PMIBSF	0.5833	0.250	0.462	0.539	0.560	0.694	0.327

3　结　论

本文基于部分互信息和贝叶斯网络互信息测试打分函数,提出一种基因调控网络构建算法(PMIBSF)。PMIBSF 算法采用 PMI 高精度构建基因相关网络,利用贝叶斯网络互信息测试打分函数确定基因间调控方向,避免了互信息、条件互信息的假阳性率和假阴性率高的问题,解决了信息论方法不能确定网络方向的局限性。计算机模拟数据和真实生物基因网络数据上的仿真实验结果,说明 PMIBSF 算法可高精度构建基因调控网络,但 PMIBSF 算法计算复杂度相对较高。如何进一步降低 PMIBSF 算法的时间复杂度使其能够构建较大规模的基因调控网络,及在互信息测试打分过程中如何缩小搜索范围,快速准确构建基因调控网络方面仍需进一步研究。

参考文献:

[1] Stuart J M, Segal E, Koller D, et al. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules[J]. Science, 2003, 302(5643): 249-255

[2] Wu J, Zhao X, Lin Z, et al. Large Scale Gene Regulatory Network Inference with a Multi-Level Strategy[J]. Molecular Biosystems, 2016, 12(2): 588-597

[3] Liu F, Zhang S W, Guo W F, et al. Inference of Gene Regulatory Network Based on Local Bayesian Networks[J]. PLos Computational Biology, 2016, 12(8): e1005024

[4] Sakamoto E, Iba H. Inferring a System of Differential Equations for a Gene Regulatory Network By Using Genetic Programming [C]//Proceedings of the 2001 Congress on Evolutionary, 2001: 720-726

[5] Huynhthu V A, Irrthum A, Wehenkel L, et al. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods [J]. Plos One, 2010, 5(9):4439-4451

[6] Shmulevich I, Dougherty E R, Kim S, et al. Probabilistic Boolean Networks: a Rule-Based Uncertainty Model for Gene Regulatory Networks[J]. Bioinformatics, 2002, 18(2): 261-274

[7] Honkela A, Girardot C, Gustafson E H, et al. Model-Based Method for Transcription Factor Target Identification with Limited Data[J]. Proceedings of the National Academy of Sciences, 2010, 107(17): 7793-7798

[8] Zhu H, Rao R S P, Zeng T, et al. Reconstructing Dynamic Gene Regulatory Networks from Sample-Based Transcriptional Data [J]. Nucleic acids research, 2012, 40(21): 10657-10667

[9] Young W C, Raftery A E, Yeung K Y. Fast Bayesian Inference for Gene Regulatory Networks Using ScanBMA[J]. BMC Systems Biology, 2014, 8(1):47-47

[10] Barzel B, Barabási A L. Network Link Prediction by Global Silencing of Indirect Correlations[J]. Nature Biotechnology, 2013, 31(8): 720-725

[11] Zhang X, Liu K, Liu Z P, et al. NARROMI: a Noise and Redundancy Reduction Technique Improves Accuracy of Gene Regulatory Network Inference[J]. Bioinformatics, 2013, 29(1): 106-113

[12] Margolin A A, Nemenman I, Basso K, et al. ARACNE: an Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context[J]. BMC Bioinformatics, 2006, 7(1): S7

[13] Zhao J, Zhou Y, Zhang X, et al. Part Mutual Information for Quantifying Direct Associations in Networks[J]. Proceedings of the National Academy of Sciences, 2016, 113(18): 5130-5135

[14] Cooper G F, Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks from Data[J]. Machine Learning, 1992, 9(4): 309-347

[15] Heckerman D, Geiger D, Chickering D M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data[J]. Machine Learning, 1995, 20(3): 197-243

[16] Schwarz G. Estimating the Dimension of a Model[J]. The Annals of Statistics, 1978, 6(2): 461-464

[17] Hansen M H, Yu B. Model Selection and the Principle of Minimum Description Length[J]. Journal of the American Statistical Association, 2001, 96(454): 746-774

[18] Lam W, Bacchus F. Learning Bayesian Belief Networks: an Approach Based on the Mdl Principle[J]. Computational Intelligence, 1994, 10(3): 269-293

[19] Basso K, Margolin A A, Stolovitzky G, et al. Reverse Engineering of Regulatory Networks in Human B Cells[J]. Nature Genetics, 2005, 37(4): 382-390

[20] Janzing D, Balduzzi D, Grosse-Wentrup M, et al. Quantifying Causal Influences[J]. The Annals of Statistics, 2013, 41(5): 2324-2358

[21] Schreiber T. Measuring Information Transfer[J]. Physical Review Letters, 2000, 85(2):461-4

[22] Campos L M. A Scoring Function for Learning Bayesian Networks Based on Mutual Information and Conditional Independence Tests[J]. Journal of Machine Learning Research, 2006, 7(2): 2149-2187

[23] Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in Silico Benchmark Generation and Performance Profiling of Network Inference Methods[J]. Bioinformatics, 2011, 27(16): 2263-2270

[24] Cantone I, Marucci L, Iorio F, et al. A Yeast Synthetic Network for in Vivo Assessment of Reverse-Engineering and Modeling Approaches[J]. Cell, 2009, 137(1): 172-181

[25] Shen-Orr S S, Milo R, Mangan S, et al. Network Motifs in the Transcriptional Regulation Network of Escherichia Coli[J]. Nature Genetics, 2002, 31(1): 64-68

[26] Wang Y, Joshi T, Zhang X S, et al. Inferring Gene Regulatory Networks from Multiple Microarray Datasets[J]. Bioinformatics, 2006, 22(19): 2413-2420

[27] Kalisch M, Bühlmann P. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm[J]. Journal of Machine Learning Research, 2007, 8: 613-636

Inferring Gene Regulatory Networks Based on Part Mutual Information and Bayesian Scoring Function

Liu Fei[1,2], Zhang Shaowu[1], Gao Hongyan[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: The inference of gene regulatory networks (GRNs) from expression data can mine the direct regulations among genes and gain deep insights into biological processes at a network level. The most widely used criteria are the Pearson correlation coefficient and partial correlation, but they can only measure linearly direct association and miss nonlinear associations. Mutual information (MI) and conditional Mutual information (CMI) not only can overcome those disadvantages, but also can process the gene expression data which are high dimensional and low samples. MI and CMI are widely used in quantifying both linear and nonlinear associations, but they suffer from the serious problems of overestimation and underestimation. GRNS based on MI and CMI suffer from higher false-positive and false-negative problem and can't identify the directions of regulatory interactions. By using the partial mutual information (PMI) and Bayesian scoring function (BSF), in this work, we present a novel algorithm (namely PMIBSF). Tested on the Synthetic networks as well as real biological molecular networks with different sizes and topologies, the results show that PMIBSF can infer RGNs with higher accuracy. The PMIBSF's performance outperforms other state-of-the-art methods, such as LP, PC-alg, NARROMI and ARACNE.

Keywords: part mutual information; mutual information test Scoring; Bayesian network; covariance matrix; gene regulatory network