Universe Scientific Publishing
Education research 教育研究
第 1 卷◆第 2 期◆版本 1.0◆2018 年 10 月
文章类型：论文|刊号（ISSN）：2630-4686

基于互信息和贝叶斯网络的改进算法流程研究

张戈辉
上海财经大学
DOI:10.18686/er.v1i2.1473

[摘　要] 基因调控网络对深入研究基因间相互作用十分重要,基于现有的基本研究方法具有的缺陷：贝叶斯网络模型不适用于大规模网络,互信息关联模型准确度不够高,将互信息关联模型和贝叶斯网络模型结合,提出了改进算法：首先利用互信息方法分割原始网络,再建立贝叶斯网络,进而对得到的网络进行聚类模块分析和 hub 基因寻找。实际试验结果显示改进算法提高了构建网络的效率,并满足了基于贝叶斯网络建立大规模网络的需求。
[关键词] 基因调控网络；贝叶斯网络；互信息

根据分子生物学的研究,基因调控在生命现象的产生过程中起着举足轻重的作用。构建基因调控网络在深入研究基因之间的相互作用以及相应的生命现象的工作中有着重要作用。在过去的几十年中,大量基因调控网络研究方法被提出,推动了生物信息学的发展。

1 材料与方法

1.1 基因调控网络

基因调控网络是指由细胞中参与基因调控作用的 DNA、RNA、蛋白质以及代谢中间物,所形成的相互作用的网络。从网络中元素的相互联系的角度,基因调控网络可以看作一个由节点(调控元素)和边(调控作用)构成的有向图。基因调控网络的本质是一个复杂的动力系统,具有复杂性、稳定性、可进化性和有限连通性。

研究基因调控网络有助于：解释基因间的相互作用,从而辅助寻找致病因子；全面认识生物大分子及其运行机制；从系统层面研究基因调控过程,进而用于分析复杂疾病发生机制。

1.2 基本的基因调控网络研究方法

1.2.1 线性组合模型

线性组合模型是一种连续的连续型网络模型,在该类模型中,一个基因的表达值通常被描述为其他几个基因表达值的加权和：

$$X_i(t+\Delta t)=\sum \omega_{ij} X_j(t) \quad (1)$$

其中,其中 Xi(t+△t)是基因 i 在时刻 t+△t 的表达水平,代表基因 j 的表达水平对基因的影响。线性组合模型求解较为复杂,有时无法得到数值解。

1.2.2 布尔网络模型

布尔网络模型是一种简单的离散型模型,网络中的每一个节点代表一个基因,每一个基因的状态用 0 或 1 来表示,0 表示处于抑制状态,1 表示处于激活状态。布尔网络模型可以用有向图 G=(V,F)表示,其中 V 表示节点集合,每个节点表示一个基因或一个环境刺激；F 表示有向边集合,每条边表示转录表达路径（如图 1）。虽然布尔网络模型较为简单,其构建网络的描述太粗糙。

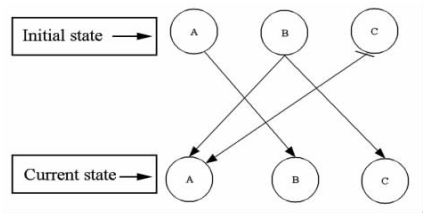
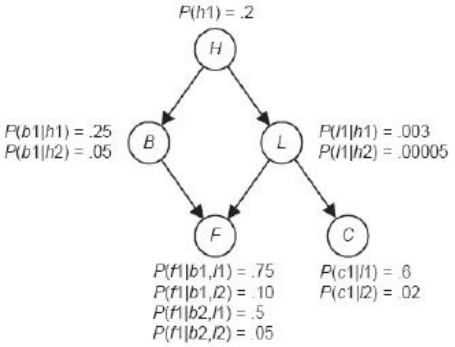


图 1　布尔网络模型

Fig.1 Boolean Network

1.2.3、贝叶斯网络模型

贝叶斯网络是一种结合了图论和概率理论的数据挖掘模型,用以表达随机变量间复杂的不确定关系。贝叶斯网络引入了有向无环图和隐形马尔可夫链来描述变量之间的相互作用,通常可以表示为 B=(G, Θ)(如图 2),其中 G 表示有向无环图,图中节点表示基因的表达向量；Θ 表示一组条件概率分布。贝叶斯网络的核心就是将条件独立关系解释为因果关系,进而解释基因调控中的因果关系。贝叶斯网络实用性强,但是并不适用于大规模网络构建。



1.2.4、互信息关联模型

互信息关联模型使用互信息描述基因与基因之间的关联,其中互信息采用信息熵作为度量。

$$H(A)=-\sum_{i=1}^{n} p(Xi)\log_2(P(Xi)) \quad (2)$$

$$H(A,B)=\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i,y_j)\log 2(\frac{1}{p(x_i,y_j)}) \quad (3)$$

$$M(A,B)=H(A)+H(B)-H(A,B) \quad (4)$$

其中,A 与 B 表示基因表达值,H(A)表示信息熵,p(Xi)表示基因表达值出现在 Xi 的频率。M(A,B)表示互信息,M(A,B)越大,A 与 B 之间相关程度越高、生物关系越密切,若 M(A,B)=0,则认为 A 与 B 之间不相关。

互信息关联模型的使用不需要前提假设,可以用于评估未知的、非线性的和其他复杂相关关系,同时互信息关联模型不依赖参数估计,因此适用于大规模的基因调控网络模型构建。但是,互信息关联模型会在一定程度上高估基因间的相关关系,并且不能分辨间接调控关系与直接调控关系,因此可能会给出过于乐观的估计。

1.3 改进现有贝叶斯网络构建算法

1.3.1 改进算法

从前文可以看到,现有方法都具有一定缺陷。但是两者不同的优势：基于模型方法的精确性和互信息关联模型的大规模适用性提供了结合两种方法以达到优势互补的思路,基于此思路可以得到以下算法(流程如图)：

(1)利用互信息算法分割原始大规模网络：分别计算所有基因两两之间的互信息,将结果由高到低排序,选择一定数目的高互信息基因对。对选出的基因对进行去重处理后得到可用基因。

(2)根据第一步得到的基因,利用爬山法(hc)在 R 语言环境中构建贝叶斯网络：从一个初始网络结构出发,通过三个搜索算子(加边、减边和转边)对当前网络结构进行修改,得到一系列候选网络结构,然后计算每个候选网络结构的评分,并选出评分最大的作为最优候选结构。如果最优候选结构的评分大于当前网络结构的评分,则以最优候选结构作为当前网络结构,继续搜索;否则,就停止搜索,并返回当前网络结构,得到需要的贝叶斯网络。

(3)基于第二步得到的贝叶斯网络进行模块聚类分析以及 hub 基因的寻找,进而进行相关的生物学意义讨论。

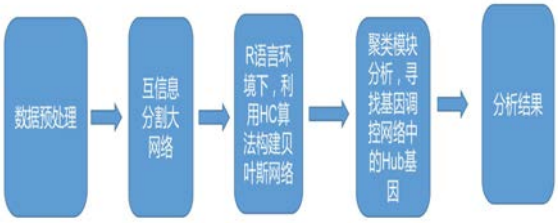


图 3　改进算法流程

Fig.3 The procedure of improved algorithm

1.3.2 改进算法使用的软件和包

(1)R 语言环境

R（https://www.r-project.org）是一种用于统计计算与绘图的语言环境。R 提供大量统计和绘图工具,并且提供开源环境。利用不同开发者提供的包,R 可以让操作者轻松的获得统计计算结果和高质量并且十分美观的图像。

(2)bnlearn 包

bnlearn（http://www.bnlearn.com）是一个应用于 R 语言环境的包,主要功能是学习贝叶斯网络的图形结构、估计贝叶斯网络的参数并做出一些有用的推断。

(3)Cytoscape 软件

Cytoscape（ http://www.cytoscape.org/index.html）是一个开源软件平台,主要用于网络和生物通路的可视化、网络与注释的整合。Cytoscape 核心部分提供基本的数据整合、分析和可视化功能,更深层的分析可以使用不同开发者提供的 Apps。

2 实验结果

2.1 基于改进算法分析玉米基因表达谱数据

实验数据来源于玉米基因表达谱,包含 28850 个基因。由于技术原因,基因表达谱数据存在缺失且单位或数量级不统一,为了结果准确,首先进行数据预处理,即缺失值处理和数据标准化。根据数据特点,这里采用数据补齐算法进行缺失值处理,并使用归一化方法将数据统一到的区间上。

完成数据预处理后,根据改进算法第一步,计算所有基因之间的互信息至少需要进行近 8.4 亿次运算,将结果降序排列后选择前 824 对基因,去重处理后得到 500 个基因。

Input	Input: The gene expression data of G
Output	Output: MI list
1	list, MI←∅
2	for each gene $ge_i \in G$
3	for each gene $ge_j \in G$
4	calculate MI(ge_i,ge_j)using Eq.(1)
5	save the MI larger than the threshold
6	end for
7	return MI list

基于这些基因,在 R 中利用 hc 算法建立贝叶斯网络。由于首先进行了大网络的分割,贝叶斯网络的建立耗时大大降低。

基于所得贝叶斯网络,利用 Cytoscape 软件完成算法第三步。首先使用 MCODE 插件进行模块聚类分析,得到 6 个显著类(如表 1)。进一步利用 cytoHubba 插件寻找 Hub 基因,得到前十个 Hub 基因(如表 2)。

表 1　模块聚类分析得到的六个显著类

Table.1 Six clusters from cluster analysis

hub 基因
GRMZM2G001200
GRMZM2G013271
GRMZM2G016805
GRMZM2G049055
GRMZM2G100913
GRMZM2G117614
GRMZM2G125342
GRMZM2G139760
GRMZM2G148913
GRMZM2G356034

表 2 玉米 hub 基因列表

Table.2 Hub gene of maize

3 讨论与展望

基于互信息关联模型以及贝叶斯网络模型的优势和劣势,本文讨论建立了将两者结合的改进算法：首先利用互信息模型分割原始大规模网络,进而建立贝叶斯网络用于基因调控网络的分析。改进算法使贝叶斯网络模型的方法可以应用于大规模网络并降低耗时。玉米基因网络的构建实验中,利用改进算法耗时较直接贝叶斯网络构建大幅减少,该实验证明了改进算法的优势以及可行性。在构建好的调控网络中,使用 Cytoscape 进行了模块分析,并找出了 10 个玉米的 hub 基因,该过程的顺利进行证明了改进算法可以为后期生物学分析提供良好的基础。

虽然改进算法的优势明显,但是仍然有一些不足之处：互信息的计算耗时耗力,同时利用互信息排序的方法分割大规模网络的准确度有待考量,例如,有些组合互信息高可能是由于两者都与第三方关系密切,目前的算法并不能识别出来这种情况。针对不足之处进一步改进算法是今后研究的重点和目标。

[参考文献]

[1]王淑栋,张善强,贺思程.基于 R 语言的互信息网络模型在乳腺癌易感基因检测分析中的应用[J].计算机系统应用,2018,27(1):143-148.

[2]金焱,胡云安,张瑾,等.互信息与爬山法相结合的贝叶斯网络结构学习[J].计算机应用与软件, 2012,29(9):122-125.

[3]王越,谭暑秋,刘亚辉.基于互信息的贝叶斯网络结构学习算法[J].计算机工程,2011,37(07):62-64.