几何模式动态贝叶斯网络推理基因调控网络

王开军，张军英，赵 峰，张宏怡
(西安电子科技大学 计算机学院，陕西 西安 710071)

摘要：针对趋势相关(两基因在其表达水平随时间上升与下降的变化趋势上相关)关系在重建基因调控网络中十分重要却尚未被挖掘利用的问题，提出了几何模式动态贝叶斯网络(Gp-DBN)方法. Gp-DBN 将每个基因的表达数据转换为一个几何模式，依据几何模式确定潜在的调控子和调控时滞，并通过推理这些几何模式之间的相关关系来发现基因间的调控关系. 该方法解决了挖掘具有趋势相关的基因调控关系的问题，能够很大程度地提高重建的基因调控网络的性能. 对 Yeast 和 E. coli 基因数据的实验结果表明无论是在无先验知识还是在有先验知识时 Gp-DBN 重建的基因调控网络的性能都比传统的动态贝叶斯网络方法有大幅度提高.

关键词：几何模式；动态贝叶斯网络；基因调控网络
中图分类号：TP18 **文献标识码**：A **文章编号**：1001-2400(2007)06-0922-04

Geometric-pattern dynamic Bayesian networks reasoning gene regulatory networks

WANG Kai-jun, *ZHANG Jun-ying*, *ZHAO Feng*, *ZHANG Hong-yi*
(School of Computer Science and Technology, Xidian Univ., Xi'an 710071, China)

Abstract: Trend correlations (i. e., two genes are correlated in their varying trends that rise and descend with time) between genes are very important but usually neglected in reconstruction of gene regulatory networks (GRN). To mine trend correlations to enhance the reconstruction performance of GRN, we propose geometric-pattern dynamic Bayesian networks (Gp-DBN). In Gp-DBN the time series of each gene is transformed to a geometric pattern, by which potential regulators and time lags are estimated, and regulatory relations between genes are discovered by reasoning correlations between these geometric patterns. Gp-DBN realizes the mining of regulatory relations with trend correlations so that it can improve the performance of GRN reconstruction. Experimental results on Yeast and E. coli data sets show that Gp-DBN improves greatly the performance of GRN reconstruction in the cases with/without prior knowledge, compared with traditional dynamic Bayesian networks.

Key Words: geometric pattern; dynamic Bayesian networks; gene regulatory networks

探索和发现基因之间的调控关系与相互作用机制是生命科学中的研究热点和前沿主题[1~4]，而从基因表达数据推理出基因调控网络[1]是其研究方向之一. 基因调控网络是一个复杂的非线性系统，从对基因间相互影响和联系的建模角度看，它是一个由节点(代表基因)和有向边(代表调控作用及方向)组成的一个有向图.

贝叶斯网络(BN)[5]是一种基于概率推理的网络模型，是由节点(代表变量)及连接节点的有向边(代表节点间的相关关系，并由条件概率表达)构成的有向无环图. 将 BN 推广到时间过程则是动态贝叶斯网络(DBN)[5,6]，而 DBN 适合于分析时间序列之间的统计相关关系而识别出调控子和它的目标基因、推理/重建基因调控网络[1]. DBN 方法已被应用于基因调控网络的重建[7~9]. 文献[9]认为考虑基因之间调控作用的时滞性(时滞性是生物学的事实[10])可以提高重建调控网络的准确性，因此提出了时滞 DBN(Lag-DBN)方法.

收稿日期：2007-06-13
基金项目：国家自然科学基金资助(60574039, 60371044)
作者简介：王开军(1965-)，男，西安电子科技大学博士研究生.

该方法将基因数据依其高低进行离散化(即若基因表达值大于其平均表达值则赋值 2,否则赋值 1),估计潜在调控子和目标基因间的调控时滞,并依据调控时滞对齐数据推理基因调控网络.这样,Lag-DBN 方法在一定程度上提高了重建调控网络的准确性.

由于基因表达水平变化的复杂性,使有调控关系的基因之间既有表达水平同时较高或较低的匹配相关性(称为同配相关)很好的情况,也有同配相关性比较差而表达水平随时间上升与下降的变化趋势的相关性(称为趋势相关)比较好的情况,并且基因表达的复杂性导致后一种情况出现得更多.现有的 DBN 方法考虑了同配相关性却忽略了趋势相关性,不能发现只具有趋势相关特性的调控关系.因此,针对挖掘和利用趋势相关关系的问题,笔者提出了几何模式动态贝叶斯网络(Gp-DBN)方法,将基因表达数据转换为基因表达的几何模式,并推理这些几何模式之间的相关关系,求得包含趋势相关关系的基因调控网络.

1　几何模式动态贝叶斯网络

挖掘基因表达的趋势相关关系可以发现具有趋势相关性的调控关系,而找到了更多的调控关系自然能提高调控网络的重建准确性.为实现此目标,需要将基因表达的时间序列转换为能描述基因表达变化趋势的几何模式,并设计推理几何模式之间相关关系的具体方法.

1.1　映射时间序列为几何模式

由随机变量的时间序列转变来的,能够表现时间序列变化趋势的几何曲线称为随机变量的几何模式.原则上讲,一个随机变量的任意复杂的时间序列可以用任意高阶多项式来拟合,但却容易导致多项式阶数过高,产生过拟合与震荡现象,不能反映时间序列的变化趋势.

这里借助现代几何的流形理论[11]来解决这个问题,将原始空间中的复杂问题分解为许多局部小问题,再将每个小问题放到一个新的类似或完全不同的空间中去求解,需要时可将求得的结果映射回原始空间.这种思想的特点是化难为易、能利用不同空间的特性和更多的数学工具.几何模式在原始空间属于可分解的复杂问题,适合于采用这种思想求解.于是,对复杂的基因时间序列,可在流形理论框架下采用分段低阶多项式曲线进行拟合,所求出的沿时间轴的二维光滑曲线 M 称为这个基因的曲线流形.这种曲线流形既能描述数据的变化趋势、又计算简单易于实现.

设系统由 q 个基因变量组成,每个基因变量有 n 个一维的连续取值的时序数据.设 $y_j \in R$ ($j=1,2,\cdots,n$) 是基因变量 Y 的时间序列.将这 n 个数据沿着时间轴 t 均匀地划分为 k 组,例如第一组有 m 个数据点 $y_j \in R$ ($j=1,2,\cdots,m$)(称为局部数据).

依据流形理论,将 M 分解成 k 个具有相同长度的曲线段(称为 M 的 k 个局部区域)来构造,每个局部区域都在相同的新空间(其坐标系称为 M 的局部坐标系)中求解(几何性质).设 M 为这 n 个数据映射成的坐标系 yot 中的二维曲线流形,M 由沿时间轴相连的 k 个局部区域组成且局部区域 V_i 对应第 i 个数据分组,则 V_i 的几何性质是由局部坐标系下用第 i 组的 m 个局部数据构造的曲线段 C_i 给出的.

设 m 个局部数据在局部坐标系 you 下表示为(u_j, y_j) ($j=1,2,\cdots,m$; $u_j=j$).考虑到拟合精度高、计算简单和避免过拟合现象的要求,采用二次多项式来拟合这 m 个局部数据,即以最小均方误差准则拟合出(u_j, y_j) 的二次多项式曲线 $C_i: y=f_i(u)=au^2+bu+c$ ($u\in[1,m]$)(解出系数 a,b,c).于是可将 C_i 的几何特性赋予 M 对应的局部区域 V_i,这称为局部同胚映射[11].求得了 k 个局部区域的几何性质则构造好了曲线流形 M,即 $\{y_j\}$ 被映射为 M:$M=\bigcup V_i$.

如上构造的局部区域 V_i 的变化趋势只与第 i 个数据组有关,使得相邻区域的变化趋势在连续性/连贯性上通常比较差.为了使相邻区域的变化趋势具有较好的连续性,以上设计再作如下改进:使用更多的局部数据构造局部曲线 C_i,即将第 i 个数据分组与相邻的两组数据(第 $i-1$ 组与 $i+1$ 组数据)合并后(共 $3m$ 个数据)构造 C_i,并且 V_i 只采用 C_i 的中间部分(对应第 i 个数据组)构造.例如,将相邻的第 1,2,3 组数据合并后构造扩展曲线段 C_2,这样 C_2 拟合了第 1,2,3 组数据,但只使用 C_2 中对应第 2 组数据的那段曲线构造局部区域 V_2.

此外,V_i 之间的连接处依据流形理论需光滑的转换以使整个 M 处处光滑(称为光滑连接).考虑到光滑连接非常复杂,上一段的连续性设计可视为对光滑连接的一种近似,而下文用到的切向量符号受这种近似连

接的影响很小,故实际中只进行近似连接.

目前采样基因数据的时间间隔较长且数据量少,需要精细的曲线流形以使得几何模式贴近数据的变化,即选取较大的分组数目 k(例如$k = n/2$),这由分组参数 p 控制:$k = n/p$. 这样求出的调控关系会比较准确,否则较小的 k 会导致曲线流形太粗泛,相应的调控关系也不准确. 由上述方法就可得到基因表达的几何模式 M,且 M 具有下文要用到的微分特性.

1.2 推理几何模式间相关关系

这里设计提取几何模式的特征、确定潜在调控子和调控时滞的方法,并采用 DBN 的推理方法求出几何模式之间联系(统计相关关系)的网络模型. 由于曲线流形 M 沿时间轴的变化趋势是由其微分特性所决定的,故可用 M 上的切向量来表示几何模式的特征. M 的参数曲线形式为$r(u) = (f(u), u)$,则其任一点处的切向量为$(f'(u), 1)$,由于切向量的变化只由 $f'(u)$ 决定,故只使用 $f'(u)$ 分量.

为了挖掘基因表达的趋势相关关系,通过计算 M 上的切向量将几何模式的特征离散表示:当某个时间点上对应的$f'(u) > 0$ 时赋值 2,当$f'(u) \leqslant 0$ 时赋值 1. 为了也挖掘同配相关关系,即发现调控子处于较高的表达水平且目标基因被调控子激活也处于较高的表达水平[12] 时的相关关系,对于 $f'(u) \leqslant 0$ 而基因表达值大于其平均表达值情况亦赋值 2. 这样的设计同时考虑了挖掘趋势相关性和同配相关性的问题. 这些以几何模式的离散特征量为主的数据将用于推理几何模式之间的关系.

进行推理之前还需要利用几何模式(离散特征量)为每个基因选择潜在的调控子和估计潜在的调控时滞. 设某个基因 g_i 的几何模式在时间 t_j 开始出现连续的$f'(u) > 0$(称为上升模式),若基因 g_1 的上升模式起始时间 t_1 比基因 g_2 的上升模式起始时间 t_2 早或相同($t_1 \leqslant t_2$),则 g_1 就被选为 g_2 的潜在调控子,并且$t_2 - t_1$ 是它们的调控时滞. 调控子 g_1 的另一个约束是 t_1 小于其时间序列总长度 T 的一半以保证 g_1 有充分的调控作用时间. 从这可以看出,利用基因的几何模式能容易地找出每个基因的潜在调控子、潜在的调控起始时间点以及调控时滞.

最后,依据调控子与目标基因间的调控时滞对齐数据(有关细节参见文献[9]),并用对齐后的数据推理基因的几何模式之间的统计相关关系. 下面讨论实现这种推理的 DBN 方法.

定义 1 设 $X^t = \{X_1^t, \cdots, X_q^t\}$ 表示 q 个离散随机变量的集合,代表时间过程在离散时间点 t 上的状态. 动态贝叶斯网络是一个通过给 $X^0, \cdots, X^n$ 指定一个概率分布 $p(X^0, \cdots, X^n \mid G, \theta)$ 来模型化时间过程的模型(G, θ),其中 G 是一个有向非循环图且 G 中的节点对应于在 X^0 和 X^1 中的随机变量,参数集合 θ 为每个节点 X_i^t 确定了一个给定父节点 $Pa(X_i^t)$ 时的条件概率分布:$p(X_i^t \mid Pa(X_i^t), G, \theta)$. $p(X^0, \cdots, X^n \mid G, \theta)$ 由因子分解公式表示[5,6]

$$p(X^0, \cdots, X^n \mid G, \theta) = \prod_{t=0}^{n} \prod_{i=1}^{q} p(X_i^t \mid Pa(X_i^t), G, \theta) \quad . \tag{1}$$

为了建模相差一定时间的时滞调控的基因网络,将如上定义扩展为广义 DBN 来使用,即将 DBN 定义中的 G 推广到非相邻的多个时间点上的随机变量,例如转移概率 $p(X^t \mid X^{t-1})$ 扩展为 $p(X^t \mid X^{t-k})$,其中 $k \in \{1, 2, \cdots, n-1\}$;而节点的联合概率分布公式(1)可以直接适用.

从给定的数据中推理 DBN 模型是指从数据中学习出网络结构 G 和对应的条件概率 $p(X_i^t \mid Pa(X_i^t))$,这是通过搜索所有可能的网络结构$\{G_i\}$ 并给每个 G_i 打分以找出最优的 DBN 结构 G_{opt} 来实现的. 常用的评判最优网络结构的评分函数是 BD 准则与 BIC 准则[13],具有最高评分(最大后验概率)的网络结构 G_{opt} 与 DBN 模型即是所求的调控网络,其中的有向边表示调控子(父节点)对目标基因(子节点)的调控关系.

若推理出几何模式之间具有统计相关关系,则其相关性必然是趋势相关性(即两个曲线流形上对应点处的切向量以高概率同为 $f'(u) > 0$ 或同为$f'(u) \leqslant 0$) 或同配相关性(即两个基因对应时间点上的表达值以高概率同时大于其平均表达值),亦 Gp-DBN 能够找出趋势相关性,也能找出同配相关性. 若某些基因之间具有这种趋势相关性而没有同配相关性,Gp-DBN 方法能够发现它们,但 Lag-DBN 方法则不能,因此 Gp-DBN 方法能找出比 Lag-DBN 方法更多的调控关系.

2 实验结果

Gp-DBN 方法给出的调控网络的结构形式与一般 DBN 方法是相同的,即有向边表示节点之间调控关系

的网络图.通常采用测量重建的基因调控网络与已知的正确调控网络的一致性程度的评价指标,即重建正确率(recall,简记为 R_c)和误识率(imprecision,简记为 I_{mp}),来评价所重建的基因调控网络的性能[14],R_c 越高而 I_{mp} 越低则重建网络与正确网络的一致性越高.设 F_N 为仅在正确网络中存在的连边数目,F_P 为仅在重建网络中存在的连边数目,T_P 为在两个网络中都存在的连边数目,则 $R_c = T_P/(T_P + F_N)$ 以及 $I_{mp} = F_P/(F_P + T_P)$[14](需本文的程序和数据,可同作者联系 sunice9@yahoo.com).

对于酵母(Yeast)细胞,文献[12]通过实验给出的基因之间的调控关系最全面,常用于检验基因调控网络的重建性能.采用文[12]中详细讨论的参与细胞周期调控的 25 个酵母基因的表达数据(alpha factor synchronization)进行实验,这些基因包括:Mbp1,Swi4,Swi6,Mcm1,Fkh1,Fkh2,Ndd1,Swi5,Ace2,Cts1,Egt2,Mcm3,Cdc46,Mcm6,Cdc6,Ste2,Ste6,Mfa2,Aga2,Hta1,Hta2,Clb2,Cdc20,Clb5,Spo12,其中前 9 个基因是调控子而其余基因是目标基因[12].第二个数据集是大肠杆菌(E. coli)的基因表达数据,其基因之间的调控关系在文献[2]中有深入的实验研究,采用文献[2]中研究过的 20 个基因的表达数据进行实验,这些基因是:fruR,tyrR,fnr,rpoE,aceA,aceB,adhE,cysG,eda,ptsH,pykF,mtr,dcuC,focA,apaH,cdsA,dapA,purA,hisJ,rfbC,其中前 4 个基因是调控子而其余基因是目标基因[2].

在实验中,时间序列转换为几何模式所采用的分组参数为 $p = 2$.在推理基因网络时,对 Yeast 数据限定最大父节点数目为 3 而对 E. coli 数据则为 2(因 E. coli 的绝大多数目标基因的调控子数目均少于 3[2]),对 Gp-DBN 和 Lag-DBN 方法均采用这个的限定.

表 1 列出了两种方法在无先验知识与有先验知识(潜在调控子的范围限定为已知的调控子)时重建基因调控网络的结果,其中"关系数目"表示找出的正确调控关系的数目.从中可以看出 Gp-DBN 方法在无和有先验知识时的正确率均比 Lag-DBN 方法高很多,同时 Gp-DBN 方法的误识率比 Lag-DBN 方法也有较大的降低.这表明挖掘基因之间的趋势相关关系确实能够有效地提高调控网络的重建性能.

表 1　两种推理方法重建基因调控网络的性能

数据集	推理方法	无先验知识			有先验知识		
		关系数目	R_c/%	I_{mp}/%	关系数目	R_c/%	I_{mp}/%
Yeast	Gp-DBN	16	59.3	73.3	20	71.4	61.5
	Lag-DBN	2	8.0	96.9	11	40.7	82.5
E. coli	Gp-DBN	9	50.0	72.7	14	77.8	53.3
	Lag-DBN	2	11.1	92.8	6	33.3	62.5

具有调控关系的基因之间通常具有同配相关性或趋势相关性或两者均存在.Gp-DBN 方法既考虑了趋势相关性也考虑了同配相关性(而 Lag-DBN 方法仅考虑了同配相关性),能发现更多的基因调控关系(参见表 1 中的关系数目).Lag-DBN 方法正确率比较低的另一个原因是:高低离散化使得某些调控子(例如 Mcm1)的二值时间序列出现交替的 1 和 2 这种震荡现象,而它的目标基因没有这种震荡现象,在求解与目标基因最相关的调控子时容易被其他连续性特征好的基因取代;而几何模式是连续性好的几何曲线,不易发生这种情况.因此 Gp-DBN 方法能获得比 Lag-DBN 方法好得多的结果,而上述实验结果也验证了这一点.

3　结 束 语

为了挖掘基因之间在表达模式的变化趋势上相关的调控关系,笔者提出了几何模式动态贝叶斯网络方法.该方法将时间序列的基因表达数据转换为几何模式,能很好地描述基因表达水平随时间上升与下降的变化趋势,并推理这些几何模式之间的相关关系来重建基因调控网络.与只挖掘同配相关关系的现有方法相比,在挖掘同配相关关系的基础上,该方法解决了挖掘趋势相关关系的问题,从而能找出更多的基因调控关系,较大幅度地提高了基因调控网络的重建性能.Yeast 和 E. coli 基因表达数据的实验结果验证了该方法的有效性.

参考文献:

[1] Friedman N. Inferring Cellular Networks Using Probabilistic Graphical Models [J]. Science, 2004, 303(5659): 799-805.

(下转第 943 页)