doi: 10.13241/j.cnki.pmb.2014.25.040

·技术与方法·

面向异步多时延基因调控网络建模的高阶动态贝叶斯网络模型及其结构学习算法*

王广云[1] 邹志康[1] 季秀才[2] 田 燕[1] 牛天慧[1]

（1 空军总医院 北京 100142；2 北京交通大学计算所 北京 100029）

摘要 目的：由基因芯片数据精确学习建模具有异步多时延表达调控关系的基因调控网络。**方法：**提出了一种高阶动态贝叶斯网络模型，并给出了网络结构学习算法，该模型假定基因的调控过程为多阶马尔科夫过程，从而能够建模基因调控网络中的异步多时延特性。**结果：**由酵母基因调控网络一个子网络人工生成了加入 10%含噪声的表达数据用于调控网络结构学习。在 75%的后验概率下，本文提出的高阶动态贝叶斯网络模型能够正确建模实际网络中全部的异步多时延调控关系，而经典动态贝叶斯网络仅能够正确建模实际网络中 1/3 的调控关系；ROC 曲线对比表明在各个后验概率水平上高阶动态贝叶斯网络模型的效果均优于经典动态贝叶斯网络。**结论：**本文提出的高阶动态贝叶斯网络模型能够精确学习建模具有异步多时延表达调控关系的基因调控网络。

关键词：基因调控网络；异步多时延；高阶动态贝叶斯网络；学习算法

中图分类号：Q-332 **文献标识码：**A **文章编号：**1673-6273(2014)25-4958-04

High-order Dynamic Bayesian Network Model and It's Structure Learning Algorithm for Constructing Gene Regulatory Networks with Asynchronous Multi-time Delays*

WANG Guang-yun[1], ZOU Zhi-kang[1], JI Xiu-cai[2], TIAN Yan[1], NIU Tian-hui[1]

(1 General Hospital of Air Force PLA, Beijing, 100142, China;

2 Institute of Computing Technology Beijing Jiaotong University, Beijing, 100029, China)

ABSTRACT Objective: To precisely construct gene regulatory networks with synchronous multi-time delays from microarray gene expression data. **Methods:** A high-order dynamic Bayesian network model and its structure learning algorithm were presented, the network model assumed that the gene regulating process was high order Markov process, so it could model the synchronous multi-time delays in gene regulation. **Results:** Artificial gene expression data with 10% noise were made from a sub-network of a yeast gene regulatory network. With 75% posterior probability, the high-order Dynamic Bayesian Network model had correctly learned all the regulatory synchronous multi-time delayed connections, while normal Dynamic Bayesian Network model had just learned 1/3 of all correct regulatory connections. The receiver operator characteristics curves showed that with any posterior probability our model was obviously much better than the normal Dynamic Bayesian Network model. **Conclusion:** The high-order Dynamic Bayesian Network can precisely model tasynchronous multi-time delays in gene regulation, and more precise gene regulation networks can be learned from microarray gene expression data by the high-order Dynamic Bayesian Network.

Key words: Gene regulatory network; Synchronous multi-time delay; High-order Dynamic Bayesian Network; Learning algorithm

Chinese Library Classification(CLC): Q-332 Document code: A

Article ID:1673-6273(2014)25-4958-04

前言

建立基因表达调控网络能够系统地考察基因间的调控关系以及调控过程，帮助人们分析基因的功能，理解生命的奥秘，同时指导医学实践[1-3]。当前，构建基因表达调控网络的主要方法之一是应用各类数据分析方法对基因芯片表达数据处理获得[4-6]。其中，贝叶斯网络模型的解释性强、灵活度高，一直以来都是基因表达调控网络构建技术研究的重点方向[7-9]。细胞中的各个基因表达调控时间并不同步，而且调控时延长度也不同[10-12]。已有的通过时序基因表达数据建立基因表达调控网络的动态贝叶斯网络模型还难以建模这种异步多时延的调控关系[7-9]。针对这一急需的问题，本文提出了一种能够对基因间的异步和多时延调控关系精确建模的高阶动态贝叶斯网络模型，可以由基因芯片时序表达数据学习获得具有异步多时延特性

* 基金项目：国家自然科学基金项目(81101177)

作者简介：王广云(1980-)，女，博士，主要研究方向：生物信息学，电话：010-66928111，E-mail: gfkdwgy@yahoo.com.cn

（收稿日期：2013-11-19 接受日期：2013-12-16）

的基因表达调控网络。

1 材料与方法

1.1 数据

本文所用数据为人工时序数据,该数据由如图 1(A)所示的酵母基因调控网络子网络[13]生成。本研究对该网络的基因进行了编号,并随机设置的调控时延,重绘了网络结构如图 1(B)所示(各边附近数字为时延长度)。根据该网络结构,本研究采用 Dirk 提出的方法生成了加入 10%随机噪声的包含 50 个采样时间点的时序基因表达数据[13]。

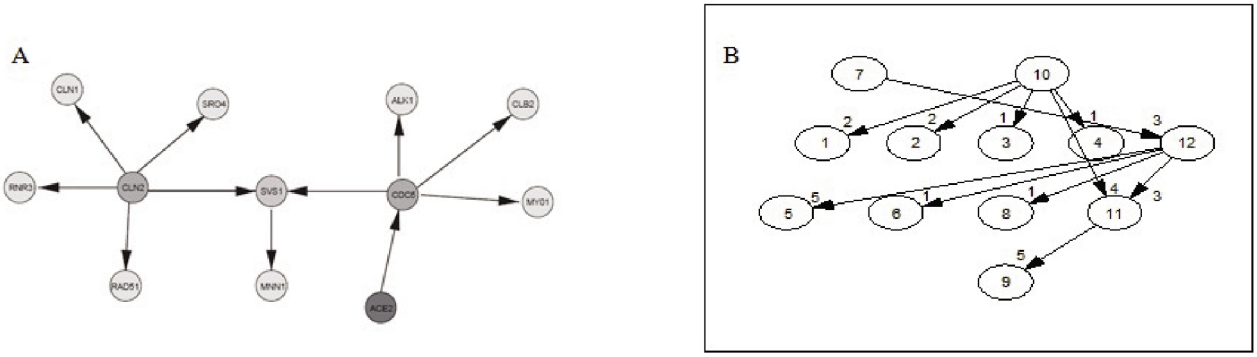

图 1 已知的酵母细胞周期基因调控网络子网络:A:原图; B:图 A 的重绘图

Fig.1 A sub- network of Yeast cell cycle gene regulatory network: A: Original drawing; B New drawing of A

1.2 方法

1.2.1 高阶动态贝叶斯网络模型 动态贝叶斯网(Dynamic Bayesian Network, DBN)是将时间信息引入到静态贝叶斯网络的随机网络模型,能够有效处理时序数据[14]。图 2 即为一个含两个节点的动态贝叶斯网络。由于动态贝叶斯网络模型能够有效描述基因间普遍存在的反馈循环调控关系和调控表达延时,所以非常适合通过基因芯片时序表达数据学习建模基因间的调控关系[15]。

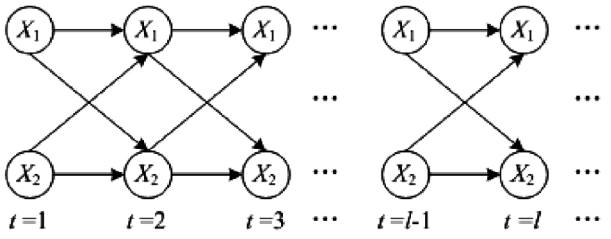

图 2 含有两个节点的动态贝叶斯网示意图

Fig.2 A Dynamic Bayesian Network with two nodes

实际上,基因间的表达调控有时并不是同步发生,不同基因的反馈循环作用周期以及不同基因之间的表达调控时延长短也很可能不同。经典动态贝叶斯网络模型无法精确建模这种异步多时延特性。为此,本文对经典动态贝叶斯网络模型进行了拓展,提出了一种高阶动态贝叶斯网络模型,称为 N 阶动态贝叶斯网络 (*N*-order Dynamic Bayesian Network, *N*-DBN),其中 N 为离散化的最大调控延时。

假设随机数据由平稳 Markov 过程产生, 作为经典的动态贝叶斯网络[15]的推广,*N* 阶动态贝叶斯网的联合概率分布可以表示为:

$$P(X_1,\cdots,X_n)\prod_{i=1}^{n}P(X_i(t)\mid X_{i\_1}(t-t_1),\cdots,X_{i\_k_i}(t-t_{k_i})) \qquad (1)$$

其中, X_{i_1}, j=1, …,k_i 为 X_i 的第 j 个父节点,k_i 为 X_i 的父结点个数, $1\leq t_j \leq N$ 为 X_i 相比于 X_{i_1} 的时延。图 3 所示的即为一个包含 3 个节点的 2 阶动态贝叶斯网络示例。显然,经典的动态贝叶斯网络为 1 阶动态贝叶斯网络。

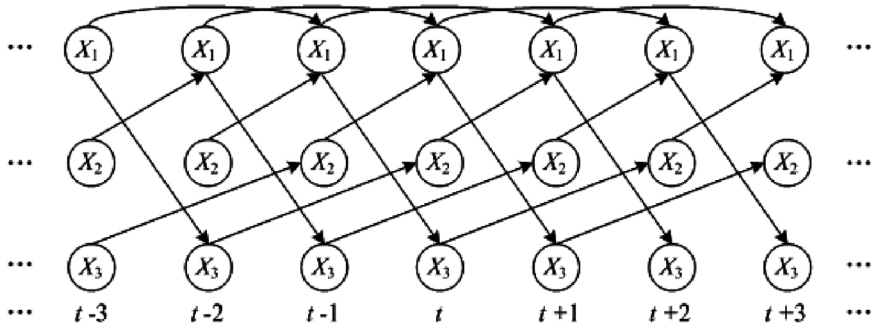

图 3 2 阶动态贝叶斯网络示例

Fig.3 Example of 2-order Dynamic Bayesian Network

1.2.2 网络结构学习算法 贝叶斯网络结构学习实际上是在网络结构空间中依据一定的度量标准搜索与网络状态观测数据集拟合得最好的网络结构的过程 [16]。本文采用较为常用的 Metropolis-Hastings 算法学习 N 阶动态贝叶斯网络结构 [17],该算法作为一种马尔科夫链蒙特卡罗(MCMC)随机采样方法,其基本思想是首先给定初始网络结构 G_0, 然后循环执行下述过程:

(1)Step1:根据建议概率 $Q(G_{mew}\mid G_{old})$和上一次循环确定

的网络 G_{old}，生成新网络 G_{new}；

（2）Step2：根据下述公式确定的接受概率随机判决是否接受新网络 G_{new}。如果选择接受 G_{new}，则将其作为网络结构的一个采样保留，并在下一次循环中用于生成新网络结构；否则，将 G_{old} 作为网络结构的采样保留，并在下一次循环中继续使用。

$$A=\min\left\{1,\frac{P(D\mid G_{new})P(G_{new})Q(G_{old}\mid G_{new})}{P(D\mid G_{old})P(G_{old})Q(G_{new}\mid G_{old})}\right\} \quad (2)$$

其中，$P(G)$为网络结构的先验概率分布，如果不存在先验知识，则有 $P(G_{new})=P(G_{old})$。$P(D\mid G)$为体现了网络结构 G 相对于数据 D 优劣的似然函数，可以作为模型结构选择的准则，一般采用评分函数（Scoring Function）计算。研究者已提出的评分函数有很多种[16]，本文所采用较为常用的 CH 评分函数，请参考文献[16]对于该函数的详细介绍。$Q(G_{new}\mid G_{old})$为建议概率，它由网络结构的建议更新方式决定。网络结构的建议更新方式一般包括增边、删边和反向三种[18]，本文采用增边和删边两种方式。通过上述算法可以得到网络结构抽样序列，可以证明该序列的极限分布为网络结构的后验概率分布 $P(G\mid D)$[17]。因此，可以由该算法产生的网络结构抽样序列估计获得最优网络结构 G^*。

2 结果

本研究对人工生成的时序数据对经典动态贝叶斯网络模型（1-DBN）与高阶动态网络模型（最大时延设置为 5，即 5-DBN）进行了对比实验验证。实验中，网络节点的入度限制为不大于 3 以保持网络结构的稀疏性。针对不同的网络模型，学习算法在不同的随机初始条件下共执行了 10 次。每次运行 2×10^5 步循环，前 1×10^5 步循环预制（Burn-in）阶段不进行的网络结构采样，后 1×10^5 步循环每间隔 100 步循环进行一次网络结构采样。因此，学习算法运行一次获得 1000 个网络结构采样。最后，我们对 10 次算法运算获得的网络结构采样集合进行统计，计算每条边的后验概率。图 4（A）和图 4（B）分别为 1-DBN 和 5-DBN 学习获得的后验概率大于 75%的边构成的调控网络。对比图 1（B），本文提出的高阶动态贝叶斯网络模型不仅构建出了真实网络的所有调控关系，而且全部正确地估计出了基因间的调控时延，而经典动态贝叶斯网络仅能够正确建模实际网络中 1/3 的调控关系。

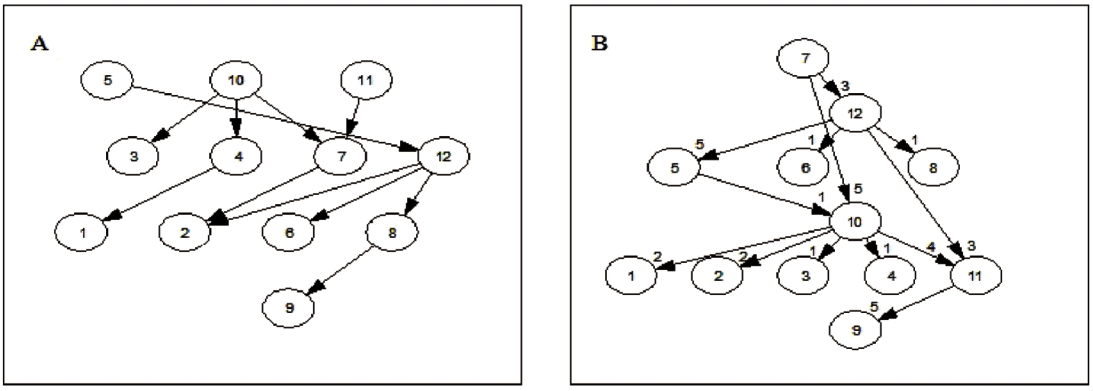

图 4 后验概率大于 75%时 1-DBN 和 5-DBN 的学习结果：A: 1-DBN; B: 5-DBN

Fig.4 The learned network structure of 1-DBN and 5-DBN with 75% with 75% posterior probability: A: 1-DBN; B: 5-DBN

为了进一步验证模型的有效性，本研究根据不同后验概率水平上 1-DBN 和 N-DBN 学习获得的网络结构进一步对比计算绘制了 1-DBN 和 N-DBN 的接受者操作特征（Receiver Operator Characteristics, ROC）曲线，如图 5 所示。ROC 曲线是一种衡量基因调控网络构建方法好坏的重要工具[19,20]。ROC 曲线中，真阳性越高（True Positive）且假阳性（False Positive）越低说明算法对于网络结构的估计结果越好。ROC 曲线对比也说明本文提出的高阶动态贝叶斯网络模型建模基因调控网络的能力明显好于经典动态贝叶斯网络模型。

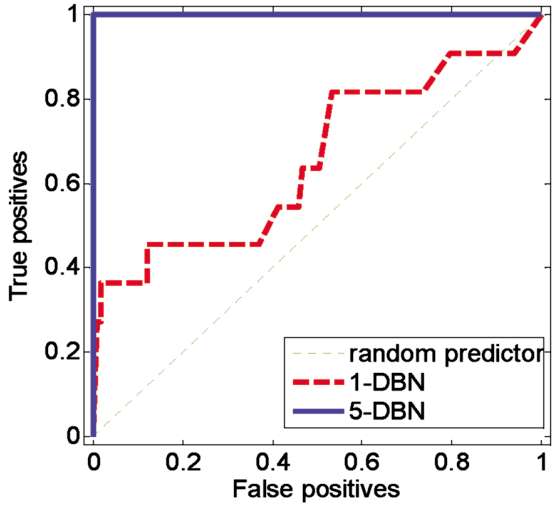

图 5 1-DBN 和 5-DBN 的 ROC 曲线

Fig.5 The ROC Curve of 1-DBN and 5-DBN

3 讨论

动态贝叶斯网络是目前由基因芯片数据构建基因调控网络主要方法。但是经典的动态贝叶斯网络难以精确建模基因表达调控中普遍存在异步多时延特性。本文对经典动态贝叶斯网络进行了推广，提出了一种高阶动态贝叶斯网络模型及结构学习算法，该高阶动态贝叶斯网络模型假定基因网络调控过程为高阶马尔科夫过程，网络模型的局部网络结构跨多个时间单位，可以精确建模异步多时延调控关系。对比实验表明相比于经典的动态贝叶斯网络模型本文提出的高阶动态贝叶斯网络模型能够更为精确地建模具有异步多时延特性的基因表达调

控网络。

应用本文提出的高阶动态贝叶斯网络模型及结构学习算法学习基因调控网络过程时需要设置网络的阶数,即基因调控网络的最大调控时延。这需要根据实际情况进行预估,我们建议以“宁大勿小”为原则估计调控网络的最大调控时延。

另外,由于实验条件、测量误差以及基因芯片的数量和质量等方面的限制,导致目前基因表达数据不够丰富且精确度不高,所以完全依靠基因表达数据构建基因调控网络的效果有时难以得到保证。为了获得更精确的基因表达调控网络,将基因芯片表达数据与已知的局部调控关系先验信息相结合是一个可行途径。后续我们将继续研究如何将基因表达调控先验信息与本文提出的高阶动态贝叶斯网络模型及结构学习算法相结合,以构建更为精确的基因表达调控网络。

参 考 文 献(References)

[1] Hinrich G, Willem T. 基于 Affymetrixi 芯片的基因表达研究[M]. 张春秀, 译. 北京: 科学出版社, 2012
Hinrich G, Willem T. Gene Expression Studies Using Affymetrixi Microarrays[M]. Zhang Chunxiu, Translating. Beijing: Science Press, 2012

[2] 李霞. 生物信息学[M]. 北京: 人民卫生出版社, 2010
Li Xia. Bioinformatics [M]. Beijing: People's Medical Publishing House, 2010

[3] 梁艳春. 生物信息学中的数据挖掘方法及应用[M]. 北京: 科学出版社, 2011
Liang Yan-chun. The Method and Application of Data Mining in Bioinformatics[M]. Beijing: Science Press, 2011

[4] Zvelebil M, Baum Jo. 理解生物信息学 [M]. 李亦学, 译. 北京: 科学出版社, 2012
Zvelebil M, Baum Jo. Understanding Bioinformatics [M]. Li Yixue, Translating. Beijing: Science Press, 2012

[5] Needham CJ, Manfield IW, Bulpitt AJ, et al. From gene expression to gene regulatory networks in Arabidopsis thaliana [J]. BMC Systems Biology, 2009, 3: 85

[6] 王明怡, 夏顺仁, 陈作舟. 基于微阵列数据的基因网络预测方法研究进展[J]. 生物物理学报. 2005, 21(1): 19-25
Wang Ming-yi, Xia Shun-ren, Chen Zuo-zhou. Progress on metods for inferring the gene networks from microarray data [J]. ACTA Biophysica Sinica, 2005, 21(1): 19-25

[7] Friedman N, Linial M, Nachman I, et al. Using Bayesian networks to analyze expression data[J]. Jounal of Computational Biology, 2000, 7 (3-4): 601-620

[8] 游顶云, 李康. 贝叶斯网络方法在基因调控研究中的应用 [J]. 中国卫生统计, 2009, 26(1): 83-86
You Ding-yun, Li Kang. The application of Bayesian network in gene regulation researches [J]. Chinese Journal of Health Statistics, 2009, 26(1): 83-86

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

[10] 徐赛娟, 郭红. 基于递归模糊神经网络的多时延基因调控网络构建方法[J]. 福州大学学报(自然科学版), 2012, 40(2): 165-171
Xu Sai-juan, Guo Hong. An approach to construct time-lagged gene regulation network based on recurrent fuzzy neural network [J]. Journal of Fuzhou University (Natural Science Edition), 2012, 40(2): 165-171

[11] 王雪松, 谷阳阳, 程玉虎. 基于复杂网络的时延基因调控网络构建[J]. 电子学报, 2010, 38(11): 2518-2522
Wang Xue-song, Gu Yang-yang, Cheng Yu-hu. Construction of Delay Gene Regulatory Network Based on Complex Network [J]. Acta Electronica Sinica, 2010, 38(11): 2518-2522

[12] 缑葵香, 宫秀军, 汤莉. 基于时序互信息构建基因调控网络 [J]. 天津大学学报, 2010, 43(7): 655-660
Gou Kui-xiang, Gong Xiu-jun, Tang Li. Constructing Gene Regulation Network Based on Time Series Mutual Information [J]. Journal of Tianjin University, 2010, 43(7): 655-660

[13] Husmeier D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks[J]. Bioinformatic, 2003, 19(17): 2271-2282

[14] Werhli AV, Husmeier D. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge [J]. Statistical Applications in Genetics and Molecular Biology, 2007, 6(1): 15

[15] Kim SY, Imoto S, Miyano S. Inferring gene networks from time series microarray data using dynamic Bayesian networks[J]. Briefings in Bioinformatics, 2003, 4(3): 228-235

[16] 王双城. 贝叶斯网络学习推理与应用 [M]. 上海: 立信会计出版社, 2010
Wang Shuang-cheng. The Learning and Application of Bayesian Network Learning[M]. Shanghai: LiXin Accounting Publishing House, 2010

[17] 赵琪. MCMC 方法研究[D]. 济南: 山东大学, 2007
Zhao Qi. Study of MCMC Method [D]. Jinan: Shandong University, 2007

[18] 岳博, 焦李成. Bayes 网络学习的 MCMC 方法[J]. 控制理论与应用, 2003, 20(4): 582-584
Yue Bo, Jiao Li-cheng. MCMC approach to Bayesian networks learning [J]. Control Theory & Applications, 2003, 20(4): 582-584

[19] 刘明辉, 王磊, 党林阁, 等. 非确定先验信息的贝叶斯网结构学习方法[J]. 计算机工程, 2010, 36(5): 165-167
Liu Ming-hui, Wang Lei, Dang Lin-ge, et al. Structure learning method of Bayesian network with uncertain prior information [J]. Computer Engineering, 2010, 36(5): 165-167

[20] 顾志峰, 叶乃好, 石耀华. 实用生物统计学 [M]. 北京: 科学出版社, 2012
Gu Zhi-feng, Ye Nai-hao, Shi Yao-hua. Practical Biostatistics [M]. Beijing: Science Press, 2012