# 基于改进 SEM 算法的基因调控网络构建方法 *

葛玲玲, 王 浩, 姚宏亮
(合肥工业大学 计算机与信息学院, 合肥 230009)

**摘 要**: 动态贝叶斯网络(DBN)是基因调控网络的一种有力建模工具。贝叶斯结构期望最大算法(SEM)能较好地处理构建基因调控网络中数据缺失的情况,但 SEM 算法学习的结果对初始参数设置依赖性强。针对此问题,提出一种改进的 SEM 算法,通过随机生成一些候选初始值,在经过一次迭代后得到的参数中选择一个最好的初始值作为模型的初始参数值,然后执行基本的 SEM 算法。利用啤酒酵母细胞周期微阵列表达数据,构建其基因调控网络并与现有文献比较,结果显示该算法进一步提高了调控网络构建的精度。

**关键词**: 基因调控网络; 动态贝叶斯网络; 贝叶斯结构期望最大化算法
**中图分类号**: TP181 **文献标志码**: A **文章编号**: 1001-3695(2010)02-0450-03
doi:10.3969/j.issn.1001-3695.2010.02.012

# Method for modeling gene regulation network based on improved structure expectation maximization algorithm

GE Ling-ling, WANG Hao, YAO Hong-liang
(*College of Computer Science & Technology, Hefei University of Technology, Hefei* 230009, *China*)

**Abstract**: Dynamic Bayesian network(DBN) is a powerful moedling tool for gene rugulation network. Missing data in building gene regulation network is better dealt with SEM(Bayesion structure expectation maximization) algorithm, however, the result of learning by SEM algorithm has strong dependence on the initial parameters. This paper proposed an improver SEM algorithm, which randomly generated a number of candidate initial parameters and selected the best parameter as whole model's initial parameter to execute basic SEM algorithm after a iterative process. Comparing gene regulation network constructed with yeast cycle gene expression data by improved SEM algorithm with existing literature, the result indicates further improve the accuracy of constructing regulation network.

**Key words**: gene rugulation network; dynamic Bayesian network(DBN); Bayesian structure expectation maximization algorithm

## 0 引言

基因调控网络的研究目的是期望从系统的角度全面揭示基因组的功能和行为。目前构建基因调控网络主要有布尔网络[1]、微分方程[2]和贝叶斯网络[3]等方法,这些方法都在不同层次上对真实的调控网络进行了抽象化。布尔网络是定性研究基因调控网络,而微分方程又是通过精细的数学分析来量化描述生物过程,但缺乏抗噪声能力,计算量大,鲁棒性能不佳。贝叶斯网络模型是这两个极端的折中。利用贝叶斯网络构建基因调控网络是目前生物信息学研究的热点[3]。

根据处理的基因表达数据类型的区别,贝叶斯网络分为静态和动态贝叶斯网络两种方法。前者适用于处理无时序信息的芯片表达数据,后者适用于处理有时序信息的数据。动态贝叶斯网络在考虑时间因素后,通过划分时间点,可以对生物过程中反馈等循环调控进行描述,克服了静态贝叶斯不能解决有向有环图的缺陷。

在构建基因调控网络的实验中所采用的微阵列数据噪声大且有很多缺失值,这种缺失值经常会对结果的精确性有影响。然而目前用于基于贝叶斯网络构建基因调控网络的贪心搜索 GS、MWST 和 K2 等网络学习算法,常常被用来处理完备数据下的结构学习[4]。Friedman 等人[5]将 EM 思想引入到丢失数据情况下的结构学习,借鉴参数学习的 EM 算法,提出贝叶斯结构期望最大(Bayesian structure expectation maximization, SEM)算法。该算法在一定程度上提高了学习效率,但是存在学习精度低、对初始参数值依赖的缺点。

本文对 SEM 算法的初始值 $\theta^0$ 的选取进行了改进,在执行基本 SEM 算法之前先随机生成一些候选初始参数,对这些初始参数分别执行 EM 算法的一次迭代,在迭代得到的值中用最大评分函数进行最优选择,把使当前网络评分最大的参数设置为整个算法的初始值 $\theta^0$,提高了算法的准确性。将此方法应用于啤酒酵母细胞周期的基因表达谱数据中,构建基因调控网络。与现有文献比较,结果表明改进后的算法进一步提高了构建调控网络的精度。

## 1 动态贝叶斯网络

动态贝叶斯网络(DBN)是建立在时间序列数据集上的一

**收稿日期**: 2009-05-31; **修回日期**: 2009-07-20 **基金项目**: 国家自然科学基金资助项目(60705015);安徽省自然科学基金资助项目(070412064);合肥工业大学科学研究发展基金资助项目(070504F)

**作者简介**:葛玲玲(1985-),女,硕士研究生,主要研究方向为贝叶斯网络学习和推理(glline@163.com);王浩(1962-),男,教授,硕导,博士,主要研究方向为人工智能、数据挖掘、软件工程等;姚宏亮(1972-),男,副教授,硕导,博士,主要研究方向为人工智能、数据挖掘.

种贝叶斯网络,把不同时间点上的随机变量区分开来,当做不同的随机变量来构建贝叶斯网络,从而避免了出现自环的情况。

从静态贝叶斯转变到动态贝叶斯,需要作一些假设和简化处理。假设条件如下:

a)在一个有限的时间内条件概率变化过程对于所有 $t$ 是一致平稳的。

b)动态概率过程是马氏的,即未来时刻的概率只与当前时刻有关而与过去时刻无关。

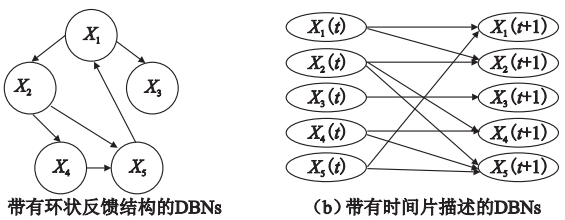
基于以上假设,建立在随机过程时间轨迹上的联合概率分布的动态贝叶斯由两部分组成:

a)一个先验网络 $B_0$。定义在初始状态 $x_0$ 上的联合概率分布。

b)一个转移网络 $B_{\rightarrow}$。定义在变量 $x_t$ 与 $x_{t+1}$ 上的转移概 $P(x_t|x_{t+1})$(对所有 $t$ 都成立)。设 $X=\{X^1,\cdots,X^n\}$ 是动态贝叶斯网络的随机变量集,$x_t^i$ 表示变量 $x^i$ 在 $t$ 时对应的随机变量,则动态贝叶斯网络在 $X=\{X^1,\cdots,X^n\}$ 上的联合概率分布可以表示为

$$P_B(X^1,\cdots,X^n)=P_{B0}(X_0)\prod_{t=0}^{T-1}P_{B_{\rightarrow}}(X_t|X_{t-1}) \tag{1}$$

静态贝叶斯网络无法描述图 1(a)中如 $X_1\rightarrow X_2\rightarrow X_5\rightarrow X_1$ 的环状反馈结构,但是在生物过程中包括很多像反馈这样的循环调控过程。动态贝叶斯网络考虑时间因素后,通过划分时间点,可以将上述反馈调控作如图 1(b)中 $X_1(t)\rightarrow X_2(t+2)\rightarrow X_5(t+3)\rightarrow X_1(t+4)$ 形式的描述。




(a) 带有环状反馈结构的DBNs　　(b) 带有时间片描述的DBNs

图1　DBNs

## 2　改进的 $SEM$ 算法

### 2.1　基本 $SEM$ 算法

SEM 算法是通过调用贝叶斯网络的推理算法填充数据集 $D$,将不完备数据下的结构学习问题转换为较容易解决的完备数据下的结构学习问题。SEM 算法主要分为结构搜索和参数学习两步。进行结构搜索时, SEM 算法使用期望充分统计因子代替不存在的充分统计因子,这样可使打分函数具有可分解形式,再进行局部搜索,以试图发现得分更高的网络结构;然后在选定的网络结构上寻找使得分最大的参数。该方法能够在一定程度上提高学习效率,并为具有缺省数据的贝叶斯结构学习提供一个框架。

其基本过程是:在每次迭代中,首先利用参数 EM 算法来实现参数最大化,然后利用 BIC 评分来进行模型选择。其中,给定初始模型 $M^0$,模型 $M^i(i>0)$ 是在 EM 算法第 $i$ 次迭代中生成,且模型序列 $M^0,\cdots,M^n$ 中的模型个数与迭代的次数相同。

算法可描述如下:

随机选定模型结构和参数 $M^0$ 及 $\theta^0$

loop $n=0,1,\cdots$,直至算法收敛

{loop $l=0,1,\cdots$,直至算法收敛或 $l=l_{\max}$

{ 使得 $\theta^{n,l+1}=\arg\max_{\theta}Q(\theta:M^n,\theta^{n,l})$;}

$\theta^{n+1}=\theta^{n,l+1}$;

将 $\mathrm{Score}(M:\theta^n,D)$ 和 $\mathrm{Score}(M:\theta^{n+1},D)$ 中得分最大的模型赋给 $M^{n+1}$;将 $M^{n+1}$ 模型的参数赋给 $\theta^{n+1,0}$}

其中:$\mathrm{Score}(M:\theta^n,D)$ 和 $\mathrm{Score}(M:\theta^{n+1},D)$ 为 BIC 评分函数;$\arg\max_{\theta}Q(\theta:M^n,\theta^{n,l})$ 为求似然函数期望最大化。

由上述算法描述可以看出,SEM 算法的执行结果对初始参数有很强的依赖性,不好的初始值会导致学习过程的循环次数增加,降低算法的时间性能和结果的学习精度。因此,怎样选取到一个好的初始参数对最后的网络结果有较大的影响。针对这个问题,本文提出了改进的 SEM 算法。

### 2.2　改进的 SEM 算法

改进的 SEM 算法是针对初始值 $\theta^0$ 的选取进行的改进,其主要思想是在执行基本的 SEM 算法之前,通过数据处理来选择出一个最佳的初始参数值,作为基本 SEM 算法的输入。获得最佳初始参数值主要包括以下几个步骤:

a)设数据集为 $D$,给定初始模型 $M^0$,在执行 SEM 算法之前,随机生成 $k$ 个初始参数值,分别记为 $\theta_1^0,\theta_2^0,\cdots,\theta_k^0$,计算当前 $\theta_i^0(i=1,\cdots,k)$ 的似然函数期望:

$$L(\theta|\theta_i^0)=\sum_L\sum_{X_L}\ln P(D_L,X_L|\theta)P(X_L|D_L,\theta_i^0) \tag{2}$$

其中:$D_L$ 和 $X_L$ 分别表示当前数据集和所有的变量。由式(2)得到 $\theta_i^0(i=1,2,\cdots,k)$,分别对应的似然函数期望,记为 $L(\theta|\theta_i^0)(i=1,2,\cdots,k)$。

b)通过最大化当前期望似然函数值,选择下一个估计:

$$\theta_i^1=\arg\max E[P(D|\theta)|D,\theta_i^0,M^0] \tag{3}$$

得到 $k$ 个结果记为 $\theta_i^1(i=1,2,\cdots,k)$。其中 arg max( )表示寻找具有最大评分的参量。设 $Q_i^0=E[P(D|\theta)|D,\theta_i^0,M^0]$,则式(3)可以写成

$$\theta_i^1=\arg\max Q_i^0(\theta,\theta_i^0) \tag{4}$$

得到 $k$ 个结果记为 $\theta_i^1(i=1,2,\cdots,k)$。

c)在这些结果中选取一个最佳值记为 $\theta^0$:

$$\theta^0=\arg\max Q(\theta_i^1)(i=1,2,\cdots,k) \tag{5}$$

d)把由式(5)得到的 $\theta^0$ 代入 SEM 算法的初始参数 $\theta^0$,执行基本的 SEM 算法。

从 $\theta_i^0$ 选择下一个估计 $\theta_i^1$ 的计算过程可以反复迭代很多次,在最不影响基本 SEM 算法时间性能的前提下,选定每个 $\theta_i^0$ 只进行一次迭代。经过这种初始化处理,可以从候选初始参数中选择出使当前网络评分得到最大值的参数作为基本算法的初始参数值,有效地减少了基本 SEM 算法执行过程中循环迭代的次数,提高了调控网络构建的精度。

## 3　实验与分析

### 3.1　实验环境和数据

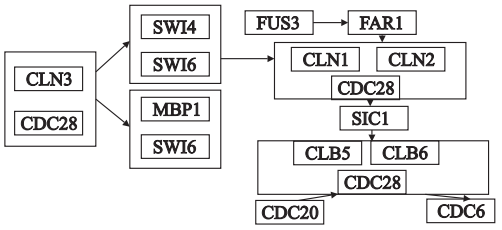
本文的实验环境 MATLAB 7.0,运行环境为操作系统 Windows XP,CPU P4 3.0 GHz,内存 512 MB,硬盘 80 GB。在网络构建实验中采用了 Leary 等人[4]编写的基于 MATLAB 的 BNT (Bayesian network toolbox) Structure Learning Package。该软件包是对 Murphy 等人开发的 BNT 工具箱在静态网络结构学习方面的扩充。

DNA 微阵列使得生物学家能够在基因组层次上研究任何种类细胞在任何时间、条件下的基因表达模式。本文使用的是由 Spellman 等人[6]于 1998 年提出的啤酒酵母 (saccharomyces

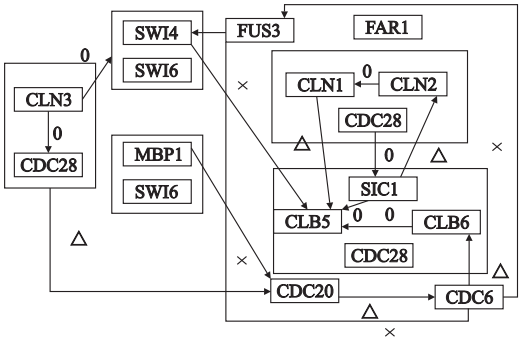
cerevisiae)细胞周期微阵列表达谱数据集。这个数据集是采用三种不同的同步化方法,最终从酵母细胞中提取出由 800 个左右表达水平符合周期变化的基因组成的,包括了这些基因在不同实验条件下或不同时间点上的具体表达水平。这些数据包含两个短的时间片(两个时间点对应实验条件为 CLN3、CLB2)和四个中等时间片(时间点分别为 18、24、17 和 14 个,对应实验条件分别为 ALPHA、CDC15、CDC28 和 ELU)。本文在基因网络的学习中,采用了四个时间片。

目标网络如图 2(a)所示,来源于 KEGG[7],是一个以 CDC28 为中心的细胞周期通路。图 2(c)表示的是 Kim 等人[8]用动态贝叶斯模型在相同的数据下构建出来的基因调控网络图。

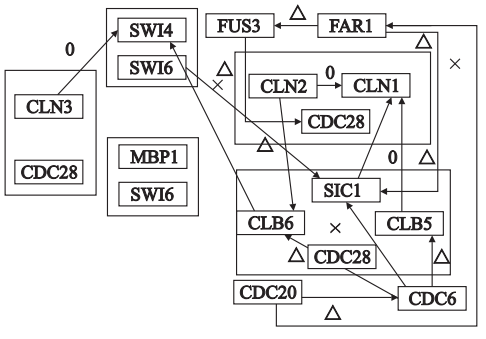
本文在实验中采用改进的 SEM 算法,分别选取了 $k=10$,15,20 构建基因调控网络,随着 $k$ 值的增大,学习的精度增加,但是同时数据处理步骤的计算量也增大。结果表明当 $k=15$ 时,算法执行的时间性能基本不变,但是学习精度明显提高,综合效率最好。构建出来的基因调控网络如图 2(b)所示。




(a) 以CDC28 为中心的细胞周期通路




(b) 本文实验得出的基因调控网络




(C) Kim等人中的实验结果

**图2　基因调控网络图**

在给出比较结果之前,首先介绍一下对实验结果的表示方法和评价标准。网络中的有向边表示两个基因之间存在调控关系,并且标志出调控与被调控的双方。笔者使用三种符号来表示预测出的调控关系是否与现有生物学资料相符。有圆圈符号的,表示已经有文献证实的调控关系;有三角符号的有向边,表示调控关系存在,但是方向与预测结果相反,或者预测的调控关系至多只跳过了一个基因,即如果有 $X_1\rightarrow X_2\rightarrow X_3$ 的调控关系存在,而算法给出的预测结果是 $X_1\rightarrow X_3$,那么这条边就会被标记为三角;如果有向边的旁边标注了叉号,则表示现有的文献并没有记载这样的调控关系。值得注意的是,这样的预测未必是错误的,而只能说现有的实验并不能够支持这样的调控关系预测而已。

### 3.2　实验结果与分析

为了定量地评价重构网络的真实性,本文引入了文献[8]中提到的 sensitivity(敏感度)和 specificity(特异度)两个指标:sensitivity = 正确估计的边数 / 目标网络中所有边数,是指网络中实际存在的调控关系中,被算法正确发现的调控关系所占的比例;specificity = 正确估计的边数/所有估计的边数,是指算法预测的调控关系总数中,正确预测所占的比例。换句话说,sensitivity 表示的是算法对于调控关系的敏感程度,specificity 衡量了算法预测的特异程度。在实验中计算 specificity 和 sensitivity 时,目标网络中总共的路径数目是 19。显然,通过本文中的方法构建出来的网络敏感度和特异度都比文献[8]中的结果好,结果表明本文提出的改进的 SEM 算法构建出来的基因调控网络更佳。本文的实验结果与文献[8]中结果的比较如表 1 所示。

表 1　本文的实验结果与文献[8]中结果的比较

| 指标 | 文献[8] | 改进的 SEM 算法 |
|---|---|---|
| 正确估计 | 4 | 6 |
| 错误估计 | 3 | 4 |
| 方向相反或者跳过一个基因数 | 8 | 5 |
| specificity | 26.7% | 40.0% |
| sensitivity | 21.1% | 31.6% |

## 4　结束语

基因调控的理论和应用已经取得了许多重大的成果,但是随着研究的不断深入,所面临的问题和挑战也越来越多。SEM 算法由于其在处理缺失值或部分可观察数据上的优势很适合从微阵列基因数据上构建基因调控网络。本文提出了一种改进的 SEM 算法,通过数据的初始化处理,从候选初始参数中选取相对最好的初始值,然后执行 SEM 算法,从而更好地构建出基因调控网络。通过实验证实了本文方法更有效,在与文献[8]的数据比较后可以看出结果网络更接近于最佳网络。但是,目前微阵列数据也存在一定的问题,如数据本身的噪声以及在离散化过程中产生出来的噪声,这些都可能会对网络的结果造成影响;同时,本文的实验中仅仅选取了节点很少的网络,而对于具体不同的网络选取什么样的 $k$ 值没有找到一个很好的评判标准,在应用到不同规模的基因调控网络构建方面有一定的局限性,需要将来进一步的完善。

**参考文献:**

[1] AKUTSU T, KUHARA S, MIYANO S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function[J]. Journal of Computational Biology, 2000, 7(3-4): 331-343.

[2] WAHDE M, HERTZ J. Coarse-grained reverse engineering of genetic regulatory networks[J]. Biosystems, 2000, 55(1-3): 129-136.

[3] FRIEDMAN N, LINIAL M, NACHMAN I, *et al*. Using Bayesian network to analyze expression data[J]. Journal of Computational Biology, 2000, 7(3-4): 601-620.

(下转第 458 页)

敛到最小点。

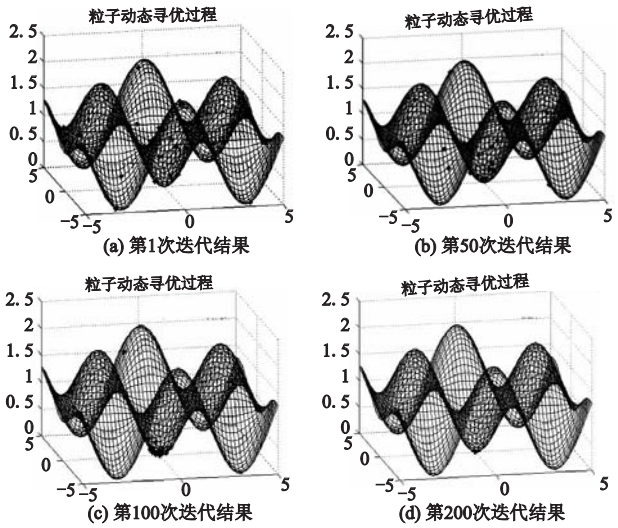



图7 二维Griewank函数动态寻优过程截图

## 5 结束语

本文分析了微粒群算法速度、种群最优值以及全局最优解之间的关系,将遗传算法的交叉和变异思想引入到粒子的位置和速度更新过程中,给出了一种解决数值优化问题的新的GAPSO算法。实验结果显示:

a)新算法不论是在单模态还是多模态 Benchmark问题上均好于现存 PSO及其改进算法,对位置更新引入的交叉思想增加了粒子的多样性和种群的进化质量,使粒子更容易找到正确的搜索方向,从而在多模态问题上易越过局部极值而向全局极值收敛;对速度变异加上判定条件,保证了变异操作不会使粒子远离全局极值,从而在单模态问题上能快速地向全局极值收敛。

b)新的 GAPSO算法提高了整个种群后期的搜索性能和最优个体的寻优能力,进一步提高了算法收敛速度,在 Shaffer问题上表现出了良好的性能。

c)交叉和变异操作的引入所增加的时间消耗在可接受的范围之内。

d)在多模态问题上 QPSO表现出的性能仅次于 GAPSO,但在单模态问题上性能却很差,其他改进算法针对不同函数表现出了各自的特点,但优化性能都不如提出的 GAPSO好。

本文不仅提出了一种新型改进 GAPSO,对模拟退火 PSO算法提出了一种新的退火方式,并对改进算法的关键参数作了深入分析,给出了 PSO算法优化的动态过程,并对目前的各种典型改进 PSO方法的仿真结果作了综合对比,选取的优化函数既有单模态又有多模态,也是优化问题的典型代表,以期对 PSO的深入研究作好前期的实验仿真基础和分析,并对以后研究 PSO算法在各种实际问题中的应用有所帮助。

**参考文献:**


[1] KENNEDY J, EBERHART R C. Particle swarm optimization[C]// Proc of IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE Press, 1995:1942-1948.

[2] PARSOPOULOS K E, VRAHATIS M N. On the computation of all global minimizers through particle swarm optimization[J]. IEEE Trans on Evolutionary Computation, 2004, 8(3):211-224.

[3] 赫然,王永吉,王青,等. 一种改进的自适应逃逸微粒群算法及实验分析[J]. 软件学报, 2005, 16(12):2036-2044.

[4] SHI Y, EBERHAART R C. A modified particle swarm optimizer [C]// Proc of Congress on Evolutionary Computation. Piscataway: IEEE Press, 1998:69-73.

[5] CLERC M. The swarm and the queen: towards a deterministic and adaptive particle swarm optimization[C]// Proc of the ICEC. [S.l.]: IEEE Press, 1999:1951-1957.

[6] 胡建秀, 曾建潮. 二阶振荡微粒群算法[J]. 系统仿真学报, 2007,19(5):997-999.

[7] SUN Jun, FENG Bin, XU Wen-bo. Particle swarm optimization with particles having quantum behavior[C]// Proc of Congress on Evolutionary Computation. 2004:325-331.

[8] 窦全胜, 周春光, 马铭. 粒子群优化的两种改进策略[J]. 计算机研究与发展, 2005,42(5):897-904.

[9] HUANG T, MOHAN A S. A hybrid boundary condition for robust particle swarm optimization[C]// Proc of IEEE Conference on Antennas and Wireless Propagation Letters. 2005:112-117.

[10] SUN JUN, XU Wen-bo, FENG Bin. Adaptive parameter control for quantum-behaved particle swarm optimization on individual level [C]// Proc of IEEE International Conference on Systems, Man and Cybernetics. 2005:3049-3054.

[11] 宋洁,董永峰,侯向丹,等. 改进的粒子群优化算法[J]. 河北工业大学学报, 2008, 37(4):55-59.

[12] 符杨,徐自力,曹家麟. 混合粒子群优化算法在电网规划中的应用[J]. 电网技术, 2008, 32(15):31-35.


(上接第 452 页)


[4] LEARY P, FRANCOIS O. BNT structure learning package: documentation and experiments[R]. 2004.

[5] FRIEDMAN N, MURPHY K, RUSSELL S. Learning the structure of dynamic probabilistic networks[C]// Proc of the 14th Conference on Uncertainty in Artificial Intelligence. 1998:139-147.

[6] SPELLMAN P T, SHERLOCK G, ZHANG M Q, et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccaromyces cerevisiae by microarray hybridization[J]. MolBiolCell, 1998, 9(12):3273-3297.

[7] Home page of KEGG[EB/OL]. http://www.genome.ad.jp/kegg.

[8] KIM S, MOTO S, MIYANO S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data[J]. Biosystems, 2004, 75(1-3): 57-65.

[9] FRIEDMAN N. The Bayesian structural EM algorithm[C]// Proc of the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kanfmann, 1998:571-578.

[10] ZHANG Yu, DENG Zhi-dong, JIANG Hong-shan, et al. Dynamic Bayesian network (DBN) with structure expectation maximization (SEM) formodeling of gene network from time series gene expression data[C]// Proc of International Conference on Bioinformatics & Computational Biology. Las Vegas, Nevada: [s.n.], 2006:26-29.

[11] 虞慧婷,吴骋,刘伟伟,等. 基因调控网络模型构建方法[J]. 第二军医大学学报, 2000, 27(7):737-740.

[12] MURPHY K, MIAN S. Modelling gene expression data using dynamic Bayesian networks[D]. Berkeley: Computer Science Division, University of California, 1999.

[13] 强波,王正志. 基于动态贝叶斯网构建基因调控网络[J]. 生物医学工程研究, 2008, 27(3):145-149.