计算机科学
COMPUTER SCIENCE

参数全局耦合的基因调控网络建模研究

马梦宇, 孙家祥, 胡春玲

引用本文

马梦宇, 孙家祥, 胡春玲. 参数全局耦合的基因调控网络建模研究[J]. 计算机科学, 2023, 50(11A): 221100088-7.

MA Mengyu, SUN Jiaxiang, HU Chunling. Modeling Gene Regulatory Networks with Global Coupling Parameters [J]. Computer Science, 2023, 50(11A): 221100088-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100088

参数全局耦合的基因调控网络建模研究

马梦宇[1]　孙家祥[1]　胡春玲[2]
1 安徽建筑大学电子与信息工程学院　合肥 230601
2 合肥学院人工智能与大数据学院　合肥 230601
(1227554288@qq.com)

摘　要　系统生物学中,基于隐马尔可夫模型的非齐次动态贝叶斯网络(HMM-DBN)能够合理推断出周期性基因表达数据中的调控关系,是重构基因调控网络的重要方法之一。但其通常假设调控参数具有完全独立性(每个时间段的参数需要独立推断),而参数假设(完全独立)等于忽略了自然界生物进化过程的连续性,这会影响网络重构精度。针对上述问题,结合多变点过程,提出了参数全局耦合的 HMM-DBN(GCHMM-DBN)。GCHMM-DBN 模型通过在 HMM-DBN 的基础上增加了全局耦合超参数,在所有时间段中共享具有相似高斯分布的噪声方差超参数和信噪比超参数,实现了回归参数的全局耦合,最终提高了基因调控网络的重构精度。在酿酒酵母(酵母)数据集和合成 RAF 数据集上进行实验,结果表明,与经典的同类型 HMM-DBN 模型相比,GCHMM-DBN 模型拥有更高的基因调控网络重构精度。

关键词:动态贝叶斯网络;基因调控网络;全局耦合;非齐次;MCMC
中图法分类号　TP311

Modeling Gene Regulatory Networks with Global Coupling Parameters

MA Mengyu[1],SUN Jiaxiang[1] and HU Chunling[2]
1 Department of Electronic Information Engineering,Anhui Jianzhu University,Hefei 230601,China
2 Department of Artificial Intelligence and Big Data,Hefei University,Hefei 230601,China

Abstract　In systems biology,the hidden Markov model non-homogeneous dynamic Bayesian network(HMM-DBN) can reasonably infer the regulatory relationships in periodic gene expression data and is one of the important methods to reconstruct gene regulatory networks. But it usually assumes complete independence of its regulatory parameters(the parameters of each time periods need to be inferred independently),and the parameter assumption(complete independence) is equivalent to ignore the continuity of biological evolutionary processes in nature,which affects the accuracy of network reconstruction. Aiming at the above problems and combining multiple changepoint processes,a hidden Markov model non-homogeneous dynamic Bayesian network with global coupling of parameters(GCHMM-DBN) is proposed. The GCHMM-DBN model achieves the global coupling of regression parameters by adding the global coupling hyperparameters,sharing the noise variance hyperparameters and signal-to-noise ratio hyperparameters of all time periods in the similarity Gaussian distribution based on the HMM-DBN,and finally improving the reconstruction accuracy of the gene regulation network. Experimental results on Saccharomyces cerevisiae(yeast) and synthetic RAF datasets show that the GCHMM-DBN model has higher accuracy of gene regulatory network reconstruction compared with the classical HMM-DBN model.

Keywords　Dynamic Bayesian networks,Gene regulatory networks,Global coupling,Non-homogeneous,Markov chain Monte Carlo

1　引言

随着系统生物学的发展,基因之间的调控关系逐渐成为研究热点之一。通过了解基因之间的转录网络[1]和蛋白质信号传递级联来研究生物体现象[2],能够最大限度地影响药物的作用与效果。由于基于多变点过程的动态贝叶斯网络(Changepoints Non-homogeneous Dynamic Bayesian Network,CPS-DBN)[3-5]可以较为准确地重构出基因调控网络的有向无环图,因此受到了科学界的广泛关注。

近年来,许多学者提出经典的 CPS-DBN 已经成为非齐次动态贝叶斯网络的流行模型[5-8],但其仍存在缺陷。由于 CPS-DBN 不能使距离较远的时间片段分配到相同的状态,导致数据的配置空间受到限制,使模型过度灵活,在其应用于短时间序列的过程中将导致过拟合。为了解决上述模型中存在的问题,学者们提出了结合经典混合模型的动态贝叶斯网络(Mixture Non-homogeneous Dynamic Bayesian Network,MIX-DBN)[9]和时间序列之间服从隐马尔可夫依赖关系的动态贝叶斯网络(Hidden Markov Model Non-homogeneous Dy-

基金项目:合肥市自然科学基金项目(2021035);国家面上基金项目(61976077)
This work was supported by the Hefei Natural Science Foundation(2021035) and National Natural Science Foundation of China(61976077).
通信作者:胡春玲(huchunling@hfuu.edu.cn)

namic Bayesian Network, HMM-DBN)[10]。其中 MIX-DBN 把数据点无限制地自由分配给不同的状态且不考虑数据点的时间顺序,HMM-DBN 在考虑数据点的时间顺序的同时,不限制数据点的分配空间,在分割周期性实验数据的过程中存在优势。因此,本文重点研究参数完全独立的 HMM-DBN 的改进方法。

在基因调控网络的学习过程中,网络结构的学习对应生物适应环境变化的细胞过程[11]。在生物适应环境的过程中,短时间内从头构建完全不同的基因调控网络是不现实的,即已经建立的基因调控关系在短期内不会改变,新的调控关系需要在已知的调控关系的基础上继续发展。因此,在学习网络结构的过程中需要考虑这种连续性过程。而 HMM-DBN 假设参数完全独立,使每个实验条件的参数需要单独推断,忽略了生物在适应环境变化的过程中,其细胞过程的连续性,因此会影响网络结构的推断精度。

针对 HMM-DBN 在学习网络结构的过程中无法考虑连续性细胞过程的问题,本文基于文献[11]提出的全局耦合方式来改进 HMM-DBN,并提出了参数全局耦合的隐马尔可夫模型动态贝叶斯网络(Global Coupling Hidden Markov Model Non-homogeneous Dynamic Bayesian Network, GCHMM-DBN)。GCHMM-DBN 的基本思想是在多变点过程的基础上,通过方差超参数与信噪比超参数来构建全局耦合超参数向量,使每个节点的参数信息在所有时间段内共享。最终使基因调控网络的学习过程具有连续性,使其更符合生物体发展与进化的自然规律,从而实现网络重构精度的提升。

本文的主要贡献包括以下 3 个方面:

1)在 HMM-DBN 的基础上添加全局耦合超参数,使回归参数在学习网络结构的过程中保持相似。Grzegorczyk 等在 2013 年的研究中认为[11]:分段之间的回归参数保持相似可以使网络结构在学习过程中保持稳定,从而在学习网络结构的过程中考虑细胞过程的连续性。

2)在 1)的基础上增加额外的层来扩展节点特定的噪声方差超参数和信噪比超参数,使参数信息在节点间共享。根据以往的研究经验[11],节点之间不同的信息共享方式会影响网络重构精度,因此本文对 3 种不同的节点信息耦合方式进行了实验,并根据网络重构精度提升最明显的方式,提出了 GCHMM-DBN。

3)在酿酒酵母数据集和合成 RAF 数据集上进行实验,结果证明,与经典的同类型的 HMM-DBN 模型相比,GCHMM-DBN 能构建更精准的网络结构。在拟南芥数据集上进行实验,结果证明,GCHMM-DBN 可以建立起更符合现实需要的基因调控网络。

2 相关工作

基因调控网络用基因 Z_i 到基因 Z_j 的有向边($Z_i \rightarrow Z_j$)代表一个调控关系。根据基因调控过程,基因 Z_i 通过编码形成一种转录因子,该转录因子可以与基因 Z_j 的启动子结合,从而开始基因 Z_j 的转录。这种生物调控过程中的结合关系不太可能在短时间内改变,因此在最新的研究中认为:不同基因调控关系中,只有调节效应强度的灵活性会存在不同。维持生物体运作的基因调节效应强度在短时间内会随时间变化而保持稳定,从而使调控关系随时间保持稳定;适应环境的基因调节效应强度需要随时间变化及时调整以适应环境(不稳定),从而在短时间内使调控关系及时变化。因此,在生物适应环境的过程中,有些调控关系的变化过程具有稳定性,有些调控关系的变化过程具有多变性。非齐次动态贝叶斯网络(Non-homogeneous Dynamic Bayesian Network, NH-DBN)在学习网络结构的过程中,应该考虑到基因之间的调控关系为了适应环境而经历的复杂变化过程,根据调节效应强度的稳定性来学习调控关系。

针对以上问题,Grzegorczyk 于 2020 年提出了一种具有边缘耦合方案的非齐次动态贝叶斯网络模型(Edge-wise Coupling Non-homogeneous Dynamic Bayesian Network, EWC NH-DBN)[12]。与强制耦合的 NH-DBN 不同,EWC NH-DBN 不强制耦合,遵循贝叶斯范式:“让数据说话”。并且,针对每个单独的边缘(边缘方向)推断对应的交互参数是否应该与前一段的参数耦合。边缘耦合方案使用非耦合和耦合的参数作为 NH-DBN 的限制条件,因此 EWC NH-DBN 可以根据基因表达数据推断出稳定与多变的调控关系,使基因调控网络的学习过程能更接近生物的细胞过程,从而提高网络的重构精度。

但是,EWC NH-DBN 在信息耦合的过程中,只能将相邻时间段的参数信息耦合。因此,在边缘耦合方案中时间段被视为不可交换的单位,这可能导致数据的配置空间受到限制,使潜在的模型过度灵活,故本文暂不探讨边缘耦合方案改进 HMM-DBN 的方法。

3 基于隐马尔可夫模型的非齐次动态贝叶斯网络

3.1 贝叶斯回归模型

假设在一个网络 $M=\{\pi_1,\cdots,\pi_N\}$ 中有 N 个节点 $g\in\{1,\cdots,N\}$,π_g 表示节点 g 的父节点集,K_g 是状态的最大值,$y_{g,h}$ 表示节点 g 在状态 $h\in\{1,\cdots,K_g\}$ 的目标向量,$\pi_{g,h}$ 为节点 g 在状态 h 的父节点集,$X_{\pi_{g,h}}$ 为节点 g 在状态 h 的父节点集的观测值矩阵。回归模型的高斯分布为:

$$P(y_{g,h}\mid X_{\pi_{g,h}},w_{g,h},\sigma_g^2)=N(y_{g,h}\mid X_{\pi_{g,h}}^T w_{g,h},\sigma_g^2 I) \tag{1}$$

其中,I 是单位矩阵,σ_g^2 是方差超参数,段特定且条件独立的回归参数 $w_{g,h}$ 的高斯先验为:

$$P(w_{g,h}\mid\sigma_g^2,\delta_g)=N(w_{g,h}\mid 0,\delta_g\sigma_g^2 I) \tag{2}$$

其中,逆信噪比超参数 δ_g^{-1} 和逆方差超参数 σ_g^{-2} 的伽马先验为:

$$\begin{aligned}P(\sigma_g^{-2}\mid A_\sigma,B_\sigma)&=Gam(\sigma_g^{-2}\mid A_\sigma,B_\sigma)\\&=\frac{[B_\sigma]^{(A_\sigma)}}{\gamma(A_\sigma))}[\sigma_g^{-2}]^{A_\sigma-1}e^{(-B_\sigma\sigma_g^{-2})}\end{aligned} \tag{3}$$

$$\begin{aligned}P(\delta_g^{-1}\mid A_\delta,B_\delta)&=Gam(\delta_g^{-1}\mid A_\delta,B_\delta)\\&=\frac{[B_\delta]^{A_\delta}}{\gamma(A_\delta)}[\delta_\delta^{-1}]^{A_\delta-1}e^{-B_\delta\delta_g^{-1}}\end{aligned} \tag{4}$$

其中,A_δ,B_δ,A_σ,B_σ 是固定超参数,逆方差超参数 σ_g^{-2} 和信噪比超参数 δ_g^{-1} 通过 Gibbs 采样更新。

3.2 基于 HMM 的互补 Markov Chain Monte Carlo(MCMC)移动

传统的多变点过程的时间片分段的过程中,段被视为不可重复访问的单元,对于短时间序列,这可能导致过度拟合,影响网络重构精度。为了能够更合理地分配周期性实验数据的时间点,文献[10]提出了两对新的互补 MCMC 移动。

3.2.1 出生移动与死亡移动

1)出生移动。在一个时间序列中随机选出状态 k,并在属于状态 k 的时间段中随机挑选一个时间点 t,将时间点 t 之后属于状态 k 的时间段分配给新的状态 j。例如,假设分配向量$[V_g(2),\cdots,V_g(10)]=[1,1,2,1,1,1,2,1,1]$,在属于状态 1 的时间段里,随机选择一个时间点 t,并将 t 之后的所有时间点分配给状态 3,得到$[V_g(2),\cdots,V_g(10)]=[1,1,2,1,1,3,2,3,3]$。

2)死亡移动。在满足一定条件的前提下,将出生移动分配给新状态 j 的时间点重新分配到从前的状态 k。例如,假设分配向量$[V_g(2),\cdots,V_g(10)]=[1,1,2,1,1,3,2,3,3]$,其中分配给状态 3 的数据点 $T_3=\{t:V_g(t)=3\}$和分配给状态 1 的数据点 $T_1=\{t:V_g(t)=1\}$不重叠($\max(T_1)<MIN(T_3)$或 $\min(T_1)>\max(T_3)$),互补死亡移动将出生移动分配给状态 3 的数据点重新分配给状态 1,从而得到$[V_g(2),\cdots,V_g(10)]=[1,1,2,1,1,1,2,1,1]$。

3.2.2 包含移动与排除移动

1)包含移动。如果一个分配给状态 k 的时间段的前后两个时间段都属于状态 j,则将这个属于状态 k 的时间段分配给状态 j。例如,假设分配向量$[V_g(2),\cdots,V_g(10)]=[1,1,2,2,1,1,2,1,1]$,因为属于状态 2 的时间序列$[V_g(3),V_g(4)]$两边的时间序列$[V_g(1),V_g(2)]$和$[V_g(5),V_g(6)]$都属于状态 1,所以包含移动把$[V_g(3),V_g(4)]$分配给状态 1,从而得到新的分配向量$[V_g(2),\cdots,V_g(10)]=[1,1,1,1,1,1,2,1,1]$。

2)排除移动。将包含移动分配给新状态 j 的时间段的子序列分配给原来的状态 k。例如$[V_g(2),\cdots,V_g(10)]=[1,1,1,1,1,1,2,1,1]$,互补排除移动在分配给状态 1 的序列$[V_g(2),\cdots,V_g(7)]$中,随机选择一个子序列分配给状态 2,且限制子序列不能包括开始时间点 $V_g(2)$和结尾时间点 $V_g(7)$,从而得到$[V_g(2),\cdots,V_g(10)]=[1,1,2,2,2,1,2,1,1]$。

算法 1 调用两对互补 MCMC 移动进行时间分段

输入:当前节点 g 的分配向量 V_g,当前节点 g 转换点的最大状态数 k_{max},当前的网络结构 M

输出:采样得到的 V_g,k_{max}

1. 基于当前的转换点个数 k_{max},计算 a_k,b_k,c_k

2. A~rand(0,1),B~rand(0,1)

IF(A<a_k) THEN

  IF(B<c_k) THEN

    出生移动 更新 V_g,k_{max}

  ELSE THEN

    死亡移动 更新 V_g,k_{max}

IF(A<b_k) THEN

  IF(B<c_k) THEN

    包含移动 更新 V_g

  ELSE THEN

    排除移动 更新 V_g

4 参数全局耦合的 GCHMM-DBN

4.1 参数全局耦合的贝叶斯回归模型

假设回归参数 $w_{g,h}$服从共轭的高斯分布,且具有全局耦合特性,其条件分布为:

$$P(w_{g,h}\mid y_{g,h},X_{\pi_g,h},m_g,\sigma_{g,h},\delta_g)=N((\lceil\delta_gC_{g,h}\rceil^{-1}+X_{\pi_g,h}X_{\pi_g,h}^{\mathrm{T}})^{-1}(\lceil\delta_gC_{g,h}\rceil^{-1}m_g+X_{\pi_g,h}y_{g,h}),\sigma_{g,h}^2*(\lceil\delta_gC_{g,h}\rceil^{-1}+X_{\pi_g,h}X_{\pi_g,h}^{\mathrm{T}})^{-1}) \tag{5}$$

如图 1 所示,m_g 是节点 g 的全局交互超参数向量,由先验 m^*、信噪比超参数 δ_g^{-1} 与方差超参数 σ_g^2 组成的条件分布求得。当 m_g 不服从概率分布(固定)时,段特定的回归参数 $w_{g,h}$条件独立(d 分离),段与段之间没有参数信息交互,此时 MCMC 采样迭代过程无法考虑生物适应环境过程的连续性,影响了网络参数的推断,网络重构精度随之下降。因此,本文设定 m_g 具有独立的概率分布(灵活),$w_{g,h}$的 d 分离丢失,段之间参数信息共享,参数全局耦合,从而使 MCMC 采样遵循生物细胞适应环境过程的连续性,网络参数推断更合理。全局交互超参数向量的条件分布为:

$$\begin{aligned}P(m_g\mid w_{g,h},\delta_g,\sigma_g^2)=N((\sum_*^{-1}+\sum_{h=1}^{K_g}\lceil\delta_gC_{g,h}^2C_{g,h}\rceil^{-1})^{-1}\cdot\\(\sum_*^{-1}m_*+\sum_{h=1}^{K_g}\lceil\delta_gC_{g,h}^2C_{g,h}\rceil^{-1}w_{g,h})\\(\sum_*^{-1}+\sum_{h=1}^{K_g}\lceil\delta_gC_{g,h}^2C_{g,h}\rceil^{-1})^{-1})\end{aligned} \tag{6}$$

其中,$\sum_*$和 m_* 是上个节点的 m_{g-1} 的正态分布中的方差和均值。

为了研究信息在段之间和节点之间的信息耦合(共享)方式对网络重构精度的影响,本文进行了以下工作。

1)假设段特定的噪声方差超参数 $\sigma_{g,h}^2$服从固定二级参数 $A_{\sigma,g,h}$,$B_{\sigma,g,h}$的伽马分布 $\sigma_{g,h}^2\sim Gam(A_{\sigma,g,h},B_{\sigma,g,h})$。基于文献[11]的研究数据进行实验,结果表明,节点的不同段之间缺乏信息耦合将导致模型过度灵活且容易过拟合,影响网络重构精度。

2)基于工作 1)的经验,本文将信噪比超参数设定为段共享,节点特定的 σ_g^2。为了研究节点与节点之间的信息交互方式,本文提出了 3 种不同的节点之间的信息耦合方案。

S1:$\sigma_g^2\sim Gam(A_\sigma,B_\sigma)$噪声方差超参数段共享,节点特定,二级超参数 B_σ 节点共享且部分不固定(见式(10))。

S2:$\sigma_g^2\sim Gam(A_{\sigma,g},B_{\sigma,g})$噪声方差超参数段共享,节点特定,二级超参数节点特定且固定($B_{\sigma,g}$固定)。

S3:$\sigma_g^2\sim Gam(A_\sigma,B_\sigma)$噪声方差超参数段共享,节点特定,二级超参数节点共享且固定(B_σ 固定)。

实验结果显示:相比 S2 和 S3,S1 的效果更好(见 5.3.1 节),因此节点之间的信息耦合可以提高网络重构精度。本文遵循方案 S1 设置信噪比超参数与方差超参数的条件分布为:

$$\begin{aligned}&P(\delta_g^{-1}\mid y_{g,V_g},w_{g,V_g},X_{\pi_g,V_g},\sigma_g^2,m_g,A_\delta,B_\delta)\\&=Gam(A_\delta+\frac{K_gk_g}{2},B_\delta+\frac{1}{2}\sum_h\frac{1}{\sigma_g^2}[w_{g,h}-m_g]^{\mathrm{T}}C_{g,h}^{-1}[w_{g,h}-m_g])\end{aligned} \tag{7}$$

$$P(\sigma_g^{-2}\mid y_{g,V_g},X_{\pi_g,V_g},\delta_g,m_g,A_\sigma,B_\sigma)=Gam\left(A_\sigma+\frac{\sum_{h=1}^{K_g}T_{g,h}}{2},B_\sigma+\frac{\sum_{h=1}^{K_g}\Delta_{g,h}^2}{2}\right) \tag{8}$$

其中,K_g 是节点 g 的段数,k_g 是父集 π_g 的基数,$T_{g,h}$为时间段 h 的时间跨度,马氏距离 $\Delta_{g,h}^2=y_{g,h}^{\mathrm{T}}(I+\delta_gX_{\pi_g,h}^{\mathrm{T}}X_{\pi_g,h})^{-1}y_{g,h}$。$A_\sigma$,$A_\delta$ 固定超参数,B_σ,B_δ 的条件分布为:

$$P(B_\sigma\mid\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2,\alpha_\sigma,\beta_\sigma,A_\sigma)=Gam\left(\alpha_\sigma+NA_\sigma,\beta_\sigma+\sum_{g=1}^{N}\frac{1}{\sigma_g^2}\right) \tag{9}$$

$$P(B_\delta\mid\delta_1,\delta_2,\cdots,\delta_N,\alpha_\delta,\beta_\delta,A_\delta)=Gam\left(\alpha_\delta+NA_\delta,\beta_\delta+\sum_{g=1}^{N}\frac{1}{\delta_g}\right) \tag{10}$$

图 1 给出了 GCHMM-DBN 模型的紧凑模型，灰色圆圈表示固定的超参数，而白色圆圈表示使用 MCMC 推断的参数和超参数。最外面的方框代表 GCHMM-DBN 模型，中间方框代表节点 $g(g=1,\cdots,N)$，内方框代表节点特定的时间段 h $(h=1,\cdots,K_g)$。

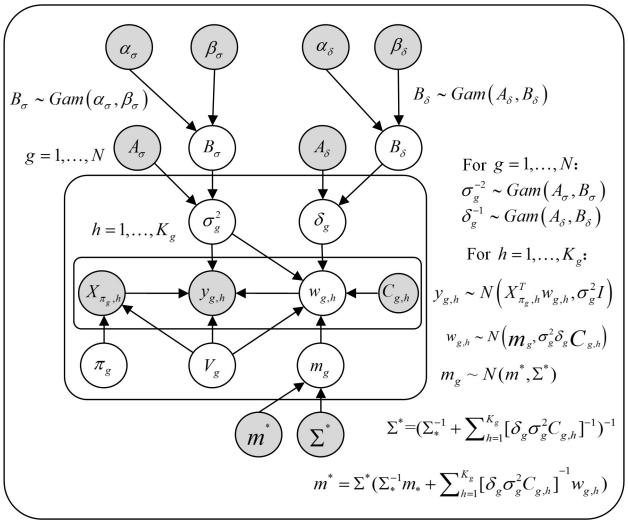


图 1 GCHMM-DBN 模型的紧凑表示

Fig. 1 Compact representation of GCHMM-DBN model

4.2 全局耦合条件下的 MCMC 采样方案

基于全局信息耦合方式，本文提出了适合的网络结构更新方法，如算法 2 所示。

算法 2 MCMC 算法的更新网络的部分伪代码

输入：信噪比超参数 $\delta_g^{(i)}$，全局交互超参数 $m_g^{(i)}$，固定二级超参数 A_σ，不固定二级超参数 $B_\delta^{(i)}$，当前父节点集 $\pi_g^{(i)}$，当前分配向量 $V_g^{(i)}$

输出：新的 $\pi_g^{(i+1)}$，新的全局交互超参数 $m_g^{(i+1)}$

1. 从后验分布 $P(\sigma_g^{-2}\mid y_{g,V_g^{(i)}},X_{\pi_g^{(i+1)},V_g^{(i)}},\delta_g^{(i)},m_g^{(i)},A_\sigma,B_\delta^{(i)})$ 中采集 $\tilde{\sigma}_g^2$。
2. 确定父节点集系统 $S(\pi_g^{(i)})$，该系统的更新方式有 3 种。1)从 $\pi_g^{(i)}$ 中删除一个父节点。2)从 $\pi_g^{(i)}$ 中添加一个父节点。3)从 $\pi_g^{(i)}$ 中替换一个父节点。从 $S(\pi_g^{(i)})$ 中随机选择一个新的候选父节点集 $\pi_g^{(*)}$。
3. 从后验分布 $P(m_g\mid\delta_g^{(i)},\tilde{\sigma}_g^2,w_{g,h})$ 采样新的全局交互超参数向量 $m_g^{(*)}$。
4. 根据条件概率式(11)决定是否使 $\pi_g^{(i)}$ 更新为 $\pi_g^{(*)}$，如果不接受移动，则 $\pi_g^{(i+1)}=\pi_g^{(i)}$，否则更新为 $\pi_g^{(*)}$。

MCMC 采样过程中得到的新父节点集合被接受的概率为：

$$A(\pi_g^{(i)}\to\pi_g^*)=MIN\left\{1,\frac{P(y_{g,v_g^{(i)}}\mid X_{\pi_g^{(*)},v_g^{(i)}},\delta_g,m_g^{(*)},A_\sigma,B_\sigma)}{P(y_{g,v_g^{(i)}}\mid X_{\pi_g^{(i)},v_g^{(i)}},\delta_g,m_g^{(8)},A_\sigma,B_\sigma)}\times\frac{|S(\pi_g^{(i)})|}{|S(\pi_g^{(*)})|}\right\}\tag{11}$$

其中，$|S(\pi_g^{(i)})|$ 为更新前父节点集系统的基数，$|S(\pi_g^{(*)})|$ 为候选父节点集系统的基数。

$$P(y_{g,V_g}\mid X_{\pi_g,V_g},\delta_g,m_g,A_\sigma,B_\sigma)=N(y_{g,h}\mid m_g,\sigma_g^2\sum_{g,v_g})\tag{12}$$

其中，$\sum_{g,V_g}=I+\delta_g X_{\pi_g,V_g}^T C_{g,V_g}X_{\pi_g,V_g}$，$m_g$ 由式(6)求得。

5 实验结果与分析

为了验证 GCHMM-DBN 比 HMM-DBN 拥有更高的网络重构精度，本文在酿酒酵母数据集、拟南芥数据集、合成 RAF 数据上进行了实验对比。实验环境为 WIN10 家庭中文版，Intel(R) Core(TM) i5-8265U CPU，64 位操作系统，20GRAM，实验软件为 matlab(2018b)。

5.1 参数设置与学习

对于初始参数值，本文设置二级超参数的初始值为：$A_\delta=2,B_\delta=0.2,A_\sigma=0.2,B_\sigma=0.01$；三级超参数的初始值为：$\alpha_\delta=0.01,\beta_\delta=2,\alpha_\sigma=20,\beta_\sigma=100$；初始全局耦合参数 $m_g=0$；初始信噪比超参数 $\delta_g=1$。

在通过表 1 完成数据的分段后，结合初始参数值通过以下步骤获得所有节点的回归参数 $w_{g,h}$ $(h=1,\cdots,K_g,g=1,\cdots,N)$：1)通过式(9)、式(10)计算二级超参数 B_σ,B_δ；2)之后通过式(7)、式(8)计算信噪比超参数 δ_g 和方差超参数 σ_g^2；3)结合式(5)和初始全局耦合参数 m_g 计算节点 g 中所有段的回归参数 $w_{g,h}$；4)结合现有的 m_g 和回归参数通过式(6)计算下一个节点 $g+1$ 的全局耦合参数 m_{g+1}；5)重复步骤 2—步骤 4，直到求出所有节点的回归参数 $w_{g,h}$。

5.2 评价标准

网络重建精度由 AUC 值来评估，对于每个网络变量 Z_j $(j=1,\cdots,N)$，得到一个后验样本 $(\pi_j^{(w)},V_j^{(w)},\sigma_j^{-2(i)},\delta_j^{(i)})_{w=1,\cdots,W}$。然后，合并采样的父节点集，形成一个图 $\{G^{(w)}\}_{w=1,\cdots,W}$ 的样本。如果 $Z_i\in\pi_j^{(w)}$，则在第 w 个图 $G^{(w)}$ 中存在边 $Z_i\to Z_j$。对于每个边 $Z_i\to Z_j$，计算边缘后验概率：

$$\hat{e}_{i,j}=\frac{1}{W}\sum_{w=1}^{W}I_{i\to j}(G^{(w)})\tag{13}$$

如果 $Z_i\in\pi_j^{(w)}$，则 $I_{i\to j}(G^{(w)})=1$，否则 $I_{i\to j}(G^{(w)})=0$。

将边缘后验概率 $\hat{e}_{i,j}$ 超过阈值 $\xi\in(0,1)$ 的所有边的集合定义为 $E(\xi)$。根据集合 $E(\xi)$ 筛选真阳性边(模型中学习出的与现实网络中对应的边)$TP(\xi)$、假阳性边(模型学习出但是不存在于现实网络中的边)$FP(\xi)$ 和假阴性边(模型没有学习出但是现实网络中存在的边)$FN(\xi)$ 的数量。本文使用两个评价指标来评估模型的网络重建能力。

式(14)—式(16)给出了评价指标的表达式。通过连接相邻点，得到精度召回率(PR)曲线。PR 曲线下面积(AUC 值)是一种通过对 PR 曲线进行数值积分来得到结果的度量方法。AUC 值或者 F_{score} 值越大，模型的网络重构能力就越强。

$$R[\xi]=TP[\xi]/(TP[\xi]+FN[\xi])\tag{14}$$

$$P[\xi]=TP[\xi]/(TP[\xi]+FP[\xi])\tag{15}$$

$$F_{score}=(2\times R[\xi]\times P[\xi])/(R[\xi]+P[\xi])\tag{16}$$

5.3 在酵母数据集上的实验结果

文献[13]综合设计了酿酒酵母中 5 个基因节点之间的基因调控网络。在 8 小时内，通过实时荧光定量 PCR 在 37 个时间点测量这些基因的表达水平，实验条件分为半乳糖/葡萄糖。真实酵母数据集的基因调控网络如图 2 所示。

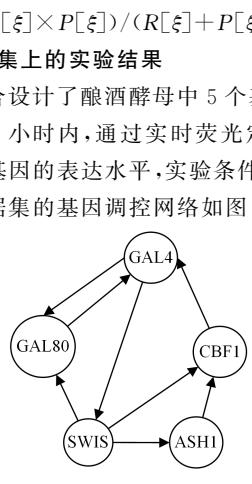


图 2 酵母基因调控网络

Fig. 2 Yeast gene regulatory network

5.3.1　不同耦合方式条件下的实验结果对比

本文的第 4.1 节提出了 3 种节点信息耦合方式：S1，S2 和 S3。图 3 给出了 3 种耦合方式对 GCHMM-DBN 的网络重构精度的提升效果。将 3 种拥有不同后验分布的方差超参数的 GCHMM-DBN 都进行 100 次单独的实验，每个实验都进行 5000 次 MCMC 迭代，取 100 次实验得到的 AUC 值的均值为最终结果。图 3 所示的实验结果显示：1）S1 比 S2 的结果更好，证明节点与节点之间有信息耦合对网络重构精度有影响；2）S2 与 S3 相比结果更好，证明每个节点之间的信息耦合强度互异。实验结果证明了 4.1 节的观点：节点间的信息交互方式会提高网络的重构精度。

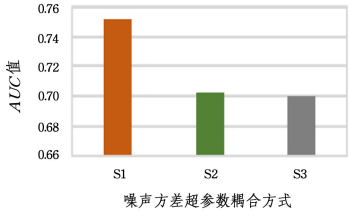


图 3　噪声方差参数耦合方式实验结果的对比

Fig. 3　Comparison of experimental results of coupling method of noise variance parameters

5.3.2　不同迭代次数条件下的实验结果对比

图 4 给出了在酵母数据集上进行实验得到的 GCHMM-DBN 和 HMM-DBN 的平均 AUC 值。实验将这两种模型在不同的 MCMC 迭代次数的前提下进行了 100 次独立的实验，将获得的 100 个 AUC 值的平均值作为最终结果。实验结果如图 4 所示，纵坐标代表平均 AUC 值，横坐标代表不同的 MCMC 迭代次数，橙色代表 GCHMM-DBN，蓝色代表 HMM-DBN。实验结果显示，与 HMM-DBN 相比，GCHMM-DBN 的平均 AUC 值提高了 1%～5%，因此 GCHMM-DBN 的网络重构精度比 HMM-DBN 更高。

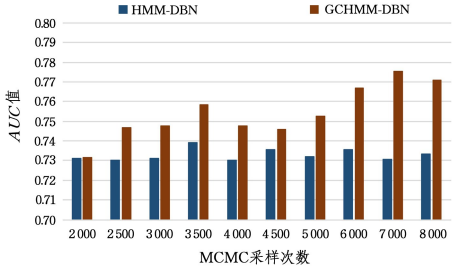


图 4　酵母数据集上不同 MCMC 迭代次数的网络重构精度对比（电子版为彩图）

Fig. 4　Comparison of network reconstruction accuracy with different MCMC iterations on yeast dataset

5.3.3　酵母数据集上的 F1 分数实验结果

图 5 给出了在酵母数据集上进行实验得到的 GCHMM-DBN 和 HMM-DBN 的 F1 分数。实验将这两种模型在 15000 次 MCMC 迭代的前提下进行了 100 次独立的实验，将获得的 100 个网络重构精度结果的平均值作为最终结果。实验结果如图 5 所示，纵坐标代表平均 F1 分数，横坐标代表不同的模型，深蓝色代表 GCHMM-DBN 的平均 F1 分数，橙色代表 HMM-DBN 的平均 F1 分数。实验结果显示，与 HMM-DBN 相比，GCHMM-DBN 的平均 F1 分数有所提高，因此 GCHMM-DBN 的网络重构能力比 HMM-DBN 更强。

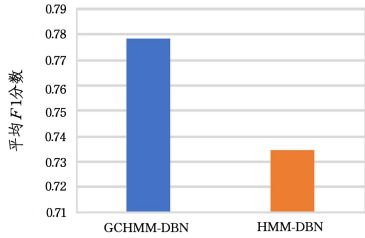


图 5　酵母数据集上求得的 F1 分数（电子版为彩图）

Fig. 5　F1 scores obtained on the yeast dataset

5.4　拟南芥网络推断

文献[14-15]提出的拟南芥的基因表达数据是从 4 个单独的实验（E1－E4）中得到的，实验条件分为有光和无光，前两个实验 E1 和 E2 中对植物进行 12:12 小时的光暗循环实验，且每隔 4 小时进行一次记录，在 E3，E4 中对植物进行 10:10 小时和 14:14 小时的光暗循环实验，每隔 4 小时进行一次记录，基因调控数据由 4 组实验数据连续排列得到。

由于拟南芥数据集没有标准的基因调控网络，无法评估网络重构精度，因此本文对 GCHMM-DBN 进行 50 000 次 MCMC 迭代，用得出的边缘后验概率来推断拟南芥基因之间的调控网络。图 6 展示的基因调控网络仅包含后验概率超过 0.5 的边，用有向边代表基因调控关系。其中被现有生物学文献描述过的边用黑色实线箭头表示，其他未被证实的调控关系用黑色虚线箭头表示。调控关系 LHY→TOC1，TOC1→LHY，ELF3→TOC1 在文献[16-18]中被证实；调控关系 ELF4→PRR9 在文献[19-20]中被证实；调控关系 ELF3→LHY 在文献[21]中被证实；调控关系 LHY→ELF4，LHY→ELF3，ELF3→PRR3 在文献[22]被证实。

与 HMM-DBN 的推测结果相比，GCHMM-DBN 推测结果比较完备，符合当下文献中列举出的大多数拟南芥基因调控关系，因此具备更高的实用价值。

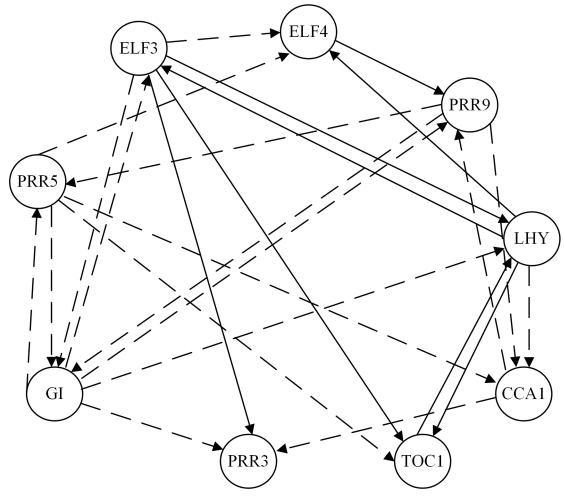


图 6　拟南芥部分基因调控网络推测结果

Fig. 6　Speculation results of some gene regulatory networks of arabidopsis thaliana

5.5　在合成 RAF 数据集上的实验结果

在合成 RAF 数据集上的实验中，本文采用的是文献[22]中论证和确定的 RAF 实验数据集，网络拓扑如图 7 所示。该网络由 11 个代表蛋白质（pip3，plcg，pip2，pkc，p38，raf，pka，

jnk, mek, erk 和 act)的节点和 20 条代表蛋白质相互作用的有向边组成。

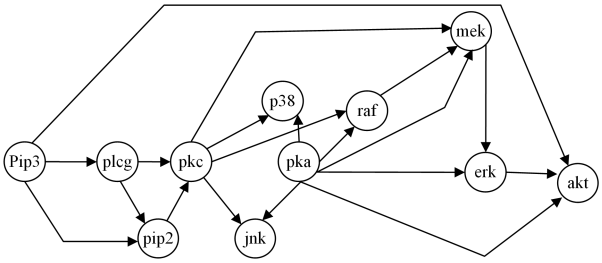


图 7 RAF 通路的拓扑结构

Fig. 7 Topology of RAF pathway

本文在合成 RAF 数据集上进行了 MCMC 迭代次数不同的多个实验，每个实验单独进行 100 次，并取 100 次实验的网络重构精度的平均值作为最终结果。实验结果如图 8 所示，图中纵坐标对应平均 AUC 值，横坐标对应的是不同迭代次数，蓝色代表 GCHMM-DBN 的 AUC 平均值，橙色代表 HMM-DBN 的 AUC 平均值。由图 8 可知，在合成 RAF 数据集上，GCHMM-DBN 的 AUC 值比 HMM-DBN 的 AUC 值高 3%～5%。

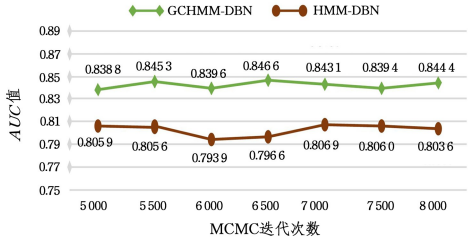


图 8 RAF 数据集上不同 MCMC 迭代次数的网络重构精度对比

(电子版为彩图)

Fig. 8 Comparison of network reconstruction accuracy with different MCMC iterations on RAF dataset

结束语 本文基于 HMM-DBN 模型，提出了新的 GCHMM-DBN，通过共享参数信息来推理网络参数。在合理分配周期性基因调控数据的基础上使基因调控网络的构建过程具有连续性，使其更符合生物体发展与进化的自然规律，从而准确地学习网络结构。GCHMM-DBN 具有以下特点：1)在 HMM-DBN 的基础上添加全局耦合超参数，使回归参数在学习网络结构的过程中保持相似；2)增加额外的层来扩展节点特定的噪声方差超参数和信噪比超参数，使参数信息在节点间共享。

本文是根据 HMM-DBN 的段之间信息不共享的缺点，结合全局耦合方法进行改进。因此，在今后的工作中希望能针对 HMM-DBN 的分段方式和全局耦合方式的细节部分进行改进：

1)在全局耦合的基础上，对马氏距离进行调整，改进 MCMC 移动的接受率。

2)在 HMM-DBN 的基础上，开发新的 MCMC 移动的方式。

参 考 文 献

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

[2] SACHS K, PEREZ O, PE'ER D, et al. Causal protein-signaling networks derived from multiparameter single-cell data[J]. Science, 2005, 308(5721): 523-529.

[3] GRZEGORCYK M, HUSMEIER D. A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology[J]. Statistical Applications in Genetics and Molecular Biology, 2012, 11(4).

[4] GRZEGORCYK M, HUSMEIER D. Bayesian Regularization of Non-homogeneous Dynamic Bayesian Networks by Globally Coupling Interaction Parameters[C]// Artificial Intelligence and Statistics. PMLR, 2012: 467-476.

[5] LEBRE S, BECQ J, DEVAUX F, et al. Statistical inference of the time-varying structure of gene-regulation networks[J]. BMC Systems Biology, 2010, 4(1): 1-16.

[6] GRZEGORCYK M, HUSMEIER D. Non-homogeneous dynamic Bayesian networks for continuous data[J]. Machine Learning, 2011, 83(3): 355-419.

[7] DONDELINGER F, LEBRE S, HUSMEIER D. Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure[J]. Machine Learning, 2013, 90(2): 191-230.

[8] DONDELINGER F, LEBRE S, HUSMEIER D. Heterogeneous continuous dynamic Bayesian networks with flexible structure and inter-time segment information sharing[C]// International Conference on Machine Learning. DBLP, 2010: 303-310.

[9] GRZEGORCYK M, HUSMEIER D. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move[J]. Machine Learning, 2008, 71(2): 265-305.

[10] GRZEGORCYKM. A non-homogeneous dynamic Bayesian network with a hidden Markov model dependency structure among the temporal data points[J]. Machine Learning, 2016, 102(2): 155-207.

[11] GRZEGORCYKM, HUSMEIER D. Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models[J]. Machine Learning, 2013, 91(1): 105-154.

[12] KAMALABAD M S, GRZEGORCYK M. A new Bayesian piecewise linear regression model for dynamic network reconstruction [J]. BMC Bioinformatics, 2021, 22(2): 1-24.

[13] 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.

[14] EDWARDS K D, ANDERSON P E, HALL A, et al. FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock[J]. The Plant Cell, 2006, 18(3): 639-650.

[15] MOCKLER T C, MICHAEL T P, PRIEST H D, et al. The DIURNAL Project: DIURNAL and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis [C]// Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory Press, 2007, 72: 353-363.

[16] HUNG F Y, CHEN F F, LI C, et al. The arabidopsis LDL1/2-

HDA6 histone modification complex is functionally associated with CCA1/LHY in regulation of circadian clock genes[J]. Nucleic acids Research,2018,46(20):10669-10681.

[17] ALABADI D,OYAMA T,YANOVSKY M J,et al. Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock[J]. Science,2001,293(5531):880-883.

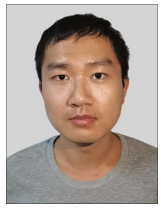
[18] KIKIS E A,KHANNA R,QUAIL P H. ELF4 is a phytochrome-regulated component of a negative-feedback loop involving the central oscillator components CCA1 and LHY[J]. The Plant Journal,2005,44(2):300-313.

[19] HERRERO E,KOLMOS E,BUJDOSO N,et al. EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the Arabidopsis circadian clock[J]. The Plant Cell,2012,24(2):428-443.

[20] ALEXANDRE MORAES T,MENGIN V,PEIXOTO B,et al. The circadian clock mutant lhy cca1 elf3 paces starch mobilization to dawn despite severely disrupted circadian clock function [J]. Plant Physiology,2022,189(4):2332-2356.

[21] KAMALABAD M S,GRZEGORCYK M. Non-homogeneous dynamic Bayesian networks with edge-wise sequentially coupled parameters[J]. Bioinformatics,2020,36(4):1198-1207.

[22] KAMALABAD M S,GRZEGORCYK M. A new Bayesian piecewise linear regression model for dynamic network reconstruction [J]. BMC Bioinformatics,2021,22(2):1-24.



MA Mengyu, born in 1998, postgraduate. His main research interests include artificial intelligence and bioinformatics.

HU Chunling, born in 1970, Ph.D, professor, M. S supervisor, is a member of China Computer Federation. Her main research interests include artificial intelligence, data mining and bioinformatics.