文章编号：2096-1472(2023)-04-24-04
DOI:10.19644/j.cnki.issn2096-1472.2023.004.006

参数边缘耦合条件下的基因调控网络建模研究

马梦宇[1]，胡春玲[2]

(1.安徽建筑大学电子与信息工程学院，安徽 合肥 230601；
2.合肥学院人工智能与大数据学院，安徽 合肥 230601)
✉1227554288@qq.com; huchunling@hfuu.edu.cn

摘 要：针对基于隐马尔科夫模型的非齐次动态贝叶斯网络(HMM-DBN)中因基因调控作用强度过度灵活而造成的网络重构精度降低的问题，提出了用参数边缘耦合方式改进HMM-DBN的方法。首先对基因调控数据进行时间分段；其次利用边缘耦合算法判断当前分段是否应该与前一段的信息交互；再次根据是否进行信息交互判断每个分段的回归参数是否耦合，结合回归参数和时间分段推断基因调控关系；最后重复上述过程直到MCMC(马尔科夫链蒙特卡洛算法)迭代完成，输出网络结构。改进后的HMM-DBN在酵母数据集与合成RAF(RAF原癌基因丝氨酸/苏氨酸-蛋白激酶)数据集上的实验结果显示，其网络重构精度达到了0.76以上，证明了该方法的有效性。

关键词：非齐次贝叶斯网络；MCMC；边缘耦合；基因调控网络
中图分类号：TP181 **文献标识码**：A

Modeling of Gene Regulatory Networks with Edge-wise Coupling Parameters

MA Mengyu[1], HU Chunling[2]

(1. *School of Electronic and Information Engineering*, *Anhui Jianzhu University*, *Hefei* 230601, *China*;
2. *Department of Artificial Intelligence and Big Data*, *Hefei University*, *Hefei* 230601, *China*)
✉1227554288@qq.com; huchunling@hfuu.edu.cn

Abstract: In order to solve the problem that the network reconstruction accuracy of Hidden Markov Model Non-homogeneous Dynamic Bayesian Network (HMM-DBN) is reduced due to the excessive flexibility of gene regulation, this paper proposes edge-wise coupling parameters to improve HMM-DBN. First, time segmentation is carried out for gene regulation data. Then, edge-wise coupling algorithm is used to judge whether the current segment should interact with the previous one. Next, whether the regression parameters of each segment are coupled is judged according to whether information interaction is conducted. The gene regulatory relationship is inferred by combining the regression parameters and time segments. Finally, the above process is repeated until MCMC (Markov Chain Monte Carlo) algorithm iteration is completed and the network structure is output. Experimental results on yeast datasets and the synthetic RAF (Raf-1 proto-oncogene, serine/threonine kinase) datasets show that the network reconstruction accuracy reaches more than 0.76, which proves the effectiveness of the proposed method.

Keywords: non-homogeneous Bayesian network; Markov Chain Monte Carlo; edge-wise coupling; generegulatory networks

1 引言(Introduction)

随着系统生物学的发展，基因调控网络逐渐成为当下生物信息学领域研究的潮流。通过了解基因之间的转录关系[1]和蛋白质信号传递级联研究生物体的基因调控网络[2]，能够有效地提升基因工程药物的作用和效果。

传统的方法是使用基于改变点过程的非齐次动态贝叶斯网络(Changepoints Non-homogeneous Dynamic Bayesian Network，CPS-DBN)[3]构建基因调控网络，CPS-DBN既能描述基因调控关系，又能描述基因调控方向，但缺点是容易导致模型过度灵活。因此，研究人员提出了具备新型分段方式的非齐次贝叶斯模型：基于隐马尔科夫模型的非齐次动态贝叶斯网络(Hidden Markov Model Non-homogeneous Dynamic Bayesian Network，HMM-DBN)[4]，HMM-DBN能将周期性实验数据中距离较远的时间点分配到相同的分

段，克服了传统CPS−DBN会导致模型过度灵活的问题。但是，由于HMM−DBN没有限制基因调控效应强度的灵活性，使基因调控关系的调控效应强度随时间推移发生较大变化，导致每个节点都要独立推断调控关系，忽略了基因调控关系为了适应环境变化可能经历的复杂过程[5]，因此影响了网络重构精度。

本文结合边缘耦合[6]的相关技术，分析了基因调控效应强度的灵活性对网络重构精度的影响，并在酵母数据集[7]和合成RAF数据集[8]上进行测试，优化了HMM−DBN，将网络重构精度提高到0.76以上。

2 边缘耦合的HMM-DBN(Edge-wise coupling HMM-DBN)

为了解决HMM−DBN中过度灵活的基因调控效应强度对学习基因调控关系的影响，进而提高网络重构精度，研究人员使用参数耦合的方式将特定参数的后验期望作为回归参数的先验分布条件，用不断迭代更新的回归参数推断不同节点之间的基因调控效应强度。通过构建耦合超参数向量，使不同的时间段之间实现信息交互，在一定程度上可以限制基因调控效应强度的灵活性，从而改善网络重构精度下降的问题。例如，顺序耦合[9]就是用前一个时间分段的回归参数的后验分布数值作为求解当前时间段的回归参数的先验分布，使回归参数随时间变化保持相似，从而让基因调控效应强度仅发生较小的变化(保持稳定)，使每个基因节点可以在已知的调控关系基础上推断自己的调控关系，最终降低了推理过程中的不确定性，使网络预测精度会得到显著的改善。但是，以上方法假设所有回归参数都随时间变化保持相似，导致调控效应强度总是保持稳定，从生物学角度来看，基因之间的调控效应强度并不会一直保持稳定，通常会受到来自变化的实验环境的影响。因此，顺序耦合不能完全模拟基因调控关系为了适应环境变化而经历的复杂变化过程，从而影响了网络重构精度。

本文根据KAMALABAD等[6]对于边缘耦合的非齐次贝叶斯网络的研究，提出了边缘耦合的基于隐马尔科夫模型的非齐次动态贝叶斯网络(Edge−wise Coupling Hidden Markov Model Non−homogeneous Dynamic Bayesian Network，EWCHMM−DBN)。EWCHMM−DBN从数据中判断当前时间段的回归参数与前一时间段的回归参数是否保持相似(耦合)，并根据实际状况在回归参数的先验分布里使用非耦合参数或耦合参数，从而区分稳定的调控效应强度和不稳定的调控效应强度。鉴于基因调控关系为了适应环境而经历的复杂变化过程，适当保留调控效应强度的灵活性可能是有用的。

2.1 边缘耦合的贝叶斯回归模型

假设一个网络中有一个节点 $g\in\{1,\cdots,N\}$，π_g 表示节点 g 的父节点集，网络结构为 $M=\{\pi_1,\cdots,\pi_N\}$，K_g 表示节点 g 中的最大数据分段数，y_h 表示在分段 $h\in\{1,\cdots,K_g\}$ 中的目标向量，X_h 为分段 h 中的观测值矩阵，第一列作为截距设置为1，$\beta_h=(\beta_{h,0},\cdots,\beta_{h,g})^{\mathrm{T}}$ 为分段 h 中所有节点的回归参数。对于每个分段 h，回归模型的高斯分布如公式(1)：

$$P(y_h\mid\beta_h,\sigma^2)\sim N(X_h\beta_h,\sigma^2 I) \tag{1}$$

其中，I 为单位矩阵。当 $h=1$时，边缘耦合方式设定回归参数 β_1 的条件分布如公式(2)：

$$P(\beta_1\mid y_1,\sigma^2,\lambda_\mu)\sim N(\tilde{\beta}_1,\sigma^2 C_1) \tag{2}$$

其中，$C_1=([\lambda_\mu I]^{-1}+X_1^T X_1)^{-1}$，$\tilde{\beta}_1=C_1X_1^T y_1$。

噪声方差超参数 σ^2 的全条件分布如公式(3)：

$$FCD(\sigma^{-2})\sim GAM(a_\sigma+0.5\cdot T,b_\sigma+0.5\cdot\Delta^2) \tag{3}$$

马氏距离 $\Delta^2:=\sum_{h=1}^{K_g}(y_h-X_h\mu_h)^T(I+X_h\Sigma_hX_h^T)^{-1}(y_h-X_h\mu_h)$，其中 $\Sigma_h:=diag\{\lambda_c\delta+\lambda_\mu(L-\delta)\}$，$\boldsymbol{L}=(1,\cdots,1)^{\mathrm{T}}$，信噪比超参数向量 $\delta=(\delta_0,\cdots,\delta_N)^{\mathrm{T}}$，$diag\{x\}$ 表示对角矩阵，当向量 x 有 n 个元素时，对角矩阵为 $n\times n$ 的矩阵，向量 x 在对角矩阵的对角线上。

因此当 $h\geqslant 2$时，β_h 的全条件分布如公式(4)：

$$FCD(\beta_h)\sim N(C_h(\Sigma_h^{-1}\mu_h+X_h^Ty_h),\sigma^2C_h) \tag{4}$$

其中，$C_h=(\Sigma_h^{-1}+X_h^TX_h)^{-1}$，$\mu_h:=\delta\cdot\tilde{\beta}_{h-1}$，$\beta_h$ 的后验期望值为 $\tilde{\beta}_h=(\Sigma_h^{-1}+X_h^TX_h)^{-1}(\Sigma_h^{-1}\mu_h+X_h^Ty_h)$。单个信噪比超参数服从伯努利分布 $\delta_i\sim BER(p)$，假设超参数 p 服从贝塔分布，$p\sim BETA(A,B)$，研究表明，当 $p\sim BETA(1,1)$，相比 $p=0.5$，并不会使网络重构精度有所提升，因此本文设定 $p=0.5$。

在以往的研究中，研究人员将基因之间的调控关系称为“边缘”，边缘分为耦合与非耦合两种。非耦合边缘对应的调控效应强度对环境影响因素敏感，回归参数随时间推移发生较大变化。耦合边缘对应的调控效应强度对环境影响因素不敏感，回归参数随时间推移保持稳定。因此在公式(4)中，当信噪比超参数 $\delta_i=0$时，代表第 i 个节点对应的数据中相邻时间段的回归参数 $\beta_{h,i}$ 与 $\beta_{h-1,i}$ 之间非耦合(不相似)，$\beta_{h,i}$ 的高斯分布为 $\beta_{h,i}=N(0,\sigma^2\lambda_\mu)$。当 $\delta_i=1$时，代表第 i 个节点对应的数据中相邻时间段的回归参数 $\beta_{h,i}$ 与 $\beta_{h-1,i}$ 之间耦合(相似)，$\beta_{h,i}$ 的高斯分布为 $\beta_{h,i}=N(\tilde{\beta}_{h-1,i},\sigma^2\lambda_c)$。

耦合参数 λ_c 与非耦合参数 λ_μ 的全条件分布如公式(5)和公式(6)：

$$FCD(\lambda_c^{-1})\sim GAM(a_c+\frac{k_c}{2},b_c+\frac{1}{2}\sigma^{-2}D_c^2) \tag{5}$$

$$FCD(\lambda_\mu^{-1})\sim GAM(a_\mu+\frac{k_\mu}{2},b_\mu+\frac{1}{2}\sigma^{-2}D_\mu^2) \tag{6}$$

其中，参数D_c^2和D_μ^2代表的数学公式如公式(7)和公式(8)：

$$D_c^2:=\sum_{h=2}^{K_g}\sum_{i:\delta_i=1}(\beta_{h,i}-\tilde{\beta}_{h-1,i})^2 \tag{7}$$

$$D_\mu^2:=\sum_{i=0}^{N}\beta_{1,i}^2+\sum_{h=2}^{K_g}\sum_{i:\delta_i=0}\beta_{h,i}^2 \tag{8}$$

其中，k_μ 是非耦合回归参数的数量，k_c 是耦合回归参数的数量，如公式(9)和公式(10)：

$$k_\mu:=(N+1)+(K_g-1)\sum_{i=0}^{N}(1-\delta_i) \tag{9}$$

$$k_c:=(K_g-1)\sum_{i=0}^{N}\delta_i \tag{10}$$

EWCHMM−DBN的图形模型如图1所示，上个时间分段的后验期望值 $\tilde{\beta}_{h-1}$ 参与当前时间分段的回归参数求解，$\tilde{\beta}_h$ 作为当前时间分段的后验期望值参与下一时间分段的回归参数求解，灰色圆圈表示固定超参数，白色圆圈表示需要MCMC采样推断的自由超参数。

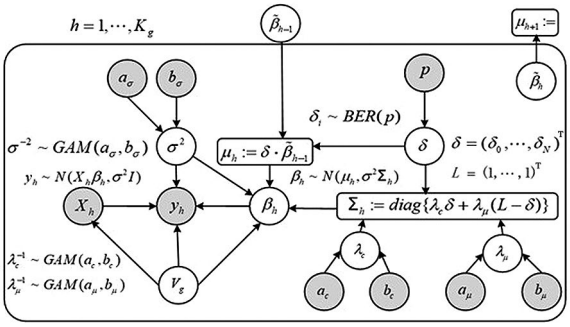


图1 EWCHMM-DBN模型的紧凑表示

Fig.1 Compact representation of the EWCHMM-DBN model

2.2 MCMC采样方案

EWCHMM-DBN将不相邻但实验条件相同的时间点分配到同一分段，例如给出9 个时间点的实验数据分配向量$\{V_g(2),\cdots,V_g(10)\}=\{1,1,2,2,1,1,2,1,1\}$，其中最大分段数$K_g=2$，不同数字代表不同分段，当多个时间点对应同一个数字时，代表这些时间点被分配到同一分段h中。EWCHMM-DBN使用包含移动、排除移动、出生移动和死亡移动对V_g进行更新，更新分配向量的方式如图2所示。

图2 时间分配向量V_g的更新过程

Fig.2 The update process of the time allocation vector V_g

每个节点g在更新分配向量V_g后，需要根据公式(11)判断是否接受此次更新，若接受，则更新V_g，不接受，则V_g保持不变。

$$P(V_g \mid K_g)=\frac{1}{K_g}\prod_{k=1}^{K_g}\frac{\gamma(\sum_{j=1}^{K_g}\alpha_{k,j})}{\gamma(\sum_{j=1}^{K_g}n_{k,j}+\alpha_{k,j})}\prod_{j=1}^{K_g}\frac{\gamma(n_{k,j}+\alpha_{k,j})}{\gamma(\alpha_{k,j})} \tag{11}$$

$n_{k,j}=|\{t \mid 3\leqslant t\leqslant T \wedge V_g(t)=j \wedge V_g(t-1)=k\}|$是序列$\{V_g(2),\cdots,V_g(10)\}$中的节点从分段$k$重新分配到分段$j$的次数，$\alpha_k=(\alpha_{k,1},\cdots,\alpha_{k,K_g})^T$为固定超参数向量。得到新的分配向量$V_g$后，更新父节点集$\pi_g^{i-1}$与父节点集集合$S(\pi_g^{(i-1)})$，$S(\pi_g^{(i-1)})$随机选择以下三种更新方式：①从$\pi_g^{i-1}$中删除一个父节点；②往$\pi_g^{i-1}$中添加一个父节点；③从$\pi_g^{i-1}$中替换一个父节点。

随后，父节点集集合$S(\pi_g^{(i-1)})$更新为$S(\pi_g^{(i)})$。从$S(\pi_g^{(i)})$中随机选择一个新的候选父节点集π_g^*，根据条件概率[公式(12)]决定是否接受π_g^*作为新的父节点集π_g^i，如果不接受则使$\pi_g^i=\pi_g^{i-1}$，反之则更新父节点集$\pi_g^i=\pi_g^*$。MCMC采样过程中得到的新父节点集被接受的概率如公式(12)：

$$A(\pi_g^{(i-1)}\rightarrow\pi_g^*)=MIN\{1,\frac{P(y_{g,h}\mid X_{\pi_g^{(*)},h},\delta_g,\lambda_\mu,\lambda_c)}{P(y_{g,h}\mid X_{\pi_g^{(i-1)},h},\delta_g,\lambda_\mu,\lambda_c)}\times \frac{P(\pi_g^{(*)})}{P(\pi_g^{(i-1)})}\times\frac{|S(\pi_g^{(i-1)})|}{|S(\pi_g^{(*)})|}\} \tag{12}$$

其中，$y_{g,h}$表示节点g在分段h的目标向量，$X_{\pi_g,h}$为节点g基于分段h和父节点集π_g的观测值矩阵，δ_g为信噪比超参数向量δ里属于当前节点的信噪比超参数。$|S(\pi_g^{(i-1)})|$为更新前的父节点集集合的基数，$|S(\pi_g^{(*)})|$为候选父节点集集合的基数，边际似然$P(y_{g,h}\mid X_{\pi_g,h},\delta_g,\lambda_\mu,\lambda_c)$如公式(13)：

$$P(y_{g,h}\mid X_{\pi_g,h},\delta_g,\lambda_\mu,\lambda_c)=\frac{\Gamma(\frac{T}{2}+a_\sigma)}{\Gamma(a_\sigma)}\cdot\frac{\pi^{-T/2}\cdot(2b_\sigma)^{a_\sigma}\cdot(2b_\sigma+\Delta^2)^{-(\frac{T}{2}+a_\sigma)}}{(\prod_{h=1}^{K_g}\det(I+X_{\pi_g,h}\Sigma_{g,h}X_{\pi_g,h}^T))} \tag{13}$$

其中，T代表总时间长度，$\Sigma_{g,h}:=diag\{\lambda_c\delta_g+\lambda_\mu(1-\delta_g)\}$。

3 实验与结果分析(Experimental results and analysis)

为了验证EWCHMM-DBN是否比HMM-DBN拥有更高的网络重构精度，本文在酵母数据集与合成RAF数据集上进行实验对比。

3.1 评价标准

本文用AUC(曲线下面积)衡量网络重构精度结果的优劣，AUC是查准率-查全率曲线下与坐标轴围成的面积，查准率-查全率曲线的文字说明如下：

对于每个网络节点$Z_j(j=1,\cdots,N)$，得到第w个后验样本($\pi_j^{(w)}$，$V_j^{(w)}$)，其中$w=1,\cdots,W$。合并采样的父节点集$\pi_j^{(w)}$，形成第w个图样本$\{G^{(w)}\}$，计算边缘后验概率如公式(14)：

$$\hat{e}_{i,j}=\frac{1}{W}\Sigma_{w=1}^{W}I_{i\rightarrow j}(G^{(w)}) \tag{14}$$

如果$Z_i\in\pi_j^{(w)}$，则在图$G^{(w)}$中存在边$Z_i\rightarrow Z_j$，$I_{i\rightarrow j}(G^{(w)})=1$，否则$I_{i\rightarrow j}(G^{(w)})=0$。

真实的网络已知，使用查准率-查全率曲线评估网络重构精度，对于每个$\psi\in[0,1]$，提取边缘后验概率$\hat{e}_{i,j}$超过ψ的边$n(\psi)$，并计算其中的真正例(真实的基因调控网络里存在的边)的数量$T(\psi)$。求得查准率$P(\psi):=T(\psi)/n(\psi)$与查全率$R(\psi):=T(\psi)/m$，并以查准率$P(\psi)$为纵坐标，查全率$R(\psi)$为横坐标，在二维坐标轴中绘制查准率-查全率曲线。其中，m是真实网络中的边数，查准率-查全率曲线下的面积称为AUC值。

3.2 在酵母数据集上的实验结果

CANTONE等[7]于2009 年综合设计了酵母基因序列中5 个基因节点之间的调控关系构成的基因调控网络，在8 h内，用实时荧光定量PCR在37 个时间节点测量了这些基因在酵母菌内部的表达水平，实验条件分为半乳糖和葡萄糖。酵母数据集中五个基因节点GAL80、GAL4、CBF1、ASH1和SWIS之间的基因调控网络如图3所示，箭头代表基因之间的调控关系。

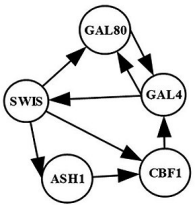


图3 酵母基因调控网络

Fig.3 Yeast gene regulatory network

图4展示了在酵母数据集上进行实验得到的EWCHMM−DBN和HMM−DBN的网络重构精度，横坐标代表不同的MCMC采样迭代次数，纵坐标代表在进行200 次独立的实验后，求出的平均 *AUC* 值。黑色代表EWCHMM−DBN的平均 *AUC* 值，灰色代表HMM−DBN的平均 *AUC* 值，如图4所示，与HMM−DBN相比，EWCHMM−DBN的平均AUC值有所提高，并达到0.76以上。

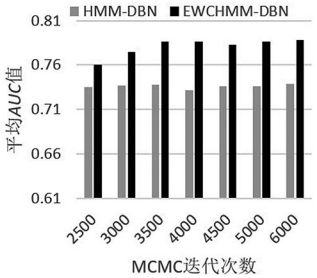
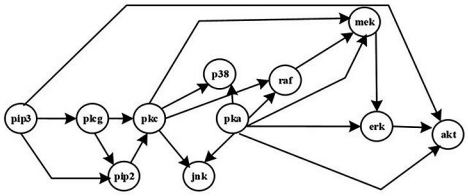


图4 酵母数据集上不同MCMC迭代次数的网络重构精度对比

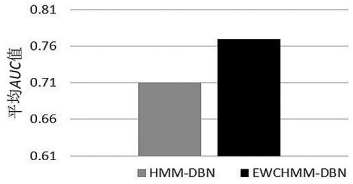
Fig.4 Comparison of network reconstruction accuracy with different MCMC iterations on yeast datasets

3.3 在合成RAF数据集上的实验结果

对于合成RAF数据集，文献[8]综合设计了实验数据，完整的网络结构如图5(a)所示，该网络由11 个节点，即pka、pip2、p38、raf、jnk、plcg、akt、erk、pip3、pkc和mek组成，有20 条代表蛋白质相互作用的有向边。图5(b)展示了在合成RAF数据集上进行实验得到的EWCHMM−DBN和HMM−DBN的平均 *AUC* 值，纵坐标对应经过200 次实验后得到的平均 *AUC* 值，横坐标对应不同的模型，黑色代表EWCHMM−DBN的平均 *AUC* 值，灰色代表HMM−DBN的平均 *AUC* 值，与HMM−DBN相比，EWCHMM−DBN的平均 *AUC* 值有所提升，并达到0.76以上。



(a)合成RAF数据集的基因调控网络



(b)在合成RAF数据集上的网络重构精度对比

图5 合成RAF数据集的网络结构与网络重构精度实验对比

Fig.5 Comparison of network structure and network reconstruction accuracy of synthetic RAF dataset

4 结论(Conclusion)

本研究使用边缘耦合的方式改进了传统的HMM−DBN，通过区分耦合与非耦合的基因调控关系，限制了基因调控效应强度的灵活性，使基因调控网络的推测过程更贴合生物适应环境的变化过程，提高了传统HMM−DBN的网络重构精度。在多个数据集上的实验结果表明：改进后的EWCHMM−DBN优于传统的HMM−DBN，证明了过度灵活的基因调控效应强度会对网络推测结果产生影响。由于影响网络重构精度的方式不止一种，因此下一步的研究计划将针对信噪比超参数和方差超参数的求解方式，尽可能地提高模型的收敛性。

参考文献(References)

[1] FRIEDMAN 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] 任洪佳.贝叶斯网络结构学习与应用研究[D].长春:吉林大学,2022.

[4] GRZEGORCZYK M. 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.

[5] GRZEGORCZYK M, 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.

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

[7] 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.

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

[9] KAMALABAD M S, GRZEGORCZYK M. Improving non−homogeneous dynamic Bayesian networks with sequentially coupled parameters[J]. Statistica Neerlandica, 2018, 72(3): 281−305.

作者简介：

马梦宇(1998−)，男，硕士生.研究领域：人工智能，生物信息学.

胡春玲(1970−)，女，博士，教授.研究领域：人工智能，数据挖掘，生物信息学.