基于复杂网络的时延基因调控网络构建

王雪松,谷阳阳,程玉虎
(中国矿业大学信息与电气工程学院,江苏徐州 221116)

摘 要: 借鉴复杂网络的分析思想和方法,采用规范化 Laplace 矩阵和 K 均值聚类法对基因调控网络进行多社团划分,同时给出每个社团内部和社团之间基因的相互作用情况.另外,为反映基因之间真实的相互作用过程和提高建模精度,在社团划分之前,采用时间序列谱分析法对基因表达时延进行精确估计.酵母细胞周期基因调控关系分析的实验结果表明,本文所提方法能更准确地反映基因之间的相互作用过程和提供基因调控模型的细节.

关键词: 复杂网络; 时延; 基因调控网络; 社团
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2010) 11-2518-05

Construction of Delay Gene Regulatory Network Based on Complex Network

WANG Xue-song, GU Yang-yang, CHENG Yu-hu
(*School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*)

Abstract: By using the analysis idea and methods of complex network, we adopted a normalized Laplace matrix and a K-mean clustering method to detect multiple communities in a gene regulatory network (GRN). At the same time, gene-gene interactions both inside a community and among communities were also described. In addition, in order to reflect the actual gene-gene interaction process and to improve the modeling precision of GRN, a time-series spectrum analysis method was used to estimate delay of gene expression before the process of community detecting. Experimental results of the cell cycle-regulated genes of yeast show that the proposed construction method of GRN not only can reflect the actual gene-gene interaction process much exactly but also can provide details about the gene regulator model.

Key words: complex network; delay; gene regulatory network; community

1 引言

基因调控网络的研究主要通过分析基因表达数据[1],结合生物信息学的方法和技术,构建合适的基因调控网络拓扑结构来模拟系统的调控机理.聚类分析是常用的构建基因调控网络的方法,它是将表达规律相似的基因聚为一类,在此基础上寻找相关基因,分析基因的功能.聚类分析有很多方法,如分级聚类[2]、K 均值聚类[3]、自组织特征映射算法[4]等.聚类算法有其显著的局限:首先聚类结果需要分离度很好的数据,几乎所有现存的算法都是从互相区别的不重叠的"类"数据中产生同样的聚类.但是,如果"类"是扩散且互相渗透,那么每种算法的结果将有所不同;其次,现有聚类方法分析的仅是简单的一对一的、线性比较关系,虽然这种方法能大大减少发现表达类型关系的计算量,但却忽视了生物系统多因素和非线性的特点;再者,聚类分析只能找到共同调控的基因,不能更精确地反映基因之间的相互作用过程和提供基因调控模型的细节.

考虑到以上聚类分析的局限性,有必要采用一种新的方法来有效划分基因聚类,同时给出每个聚类内部和聚类之间的连接情况.近年来,复杂网络在揭示各种复杂的技术网络和社会网络的形成和演化方面已经取得了一些重要进展[5,6],在生物网络的建模和分析方面也有突出成果.目前已有研究多利用复杂网络理论分析蛋白质相互作用网络,主要从复杂网络的连通性、节点的度分布、聚合系数等拓扑特征出发,使用数据挖掘和统计分析等方法进行研究[7].相比之下,复杂网络应用于基因调控网络的研究却较少.因此,有必要发展复杂网络构建基因调控网络的新方法,从而深入挖掘基因调控网络系统结构的拓扑特征,揭示基因调控网络结构与功

收稿日期:2009-07-14;修回日期:2010-01-08
基金项目:国家自然科学基金(No.60804022,No.60974050,No.61072094);教育部新世纪优秀人才支持计划(No.NCET-08-0836);江苏省自然科学基金(No.BK2008126);霍英东青年教师基金(No.121066);中国博士后科学基金特别资助(No.200902533)

能的关系.另外,现有研究在建立基因调控网络模型时,通常假设时延为0或是一个常量.但是研究发现,不同基因之间存在的调控时延是不同的.因此,有必要寻求一种比较精确的办法求解基因之间的表达时延.结合以上分析,提出一种基于复杂网络的时延基因调控网络构建方法.首先,采用谱分析方法求解基因表达时延,建立基因相互作用邻接矩阵;然后,进行复杂网络的社团划分,求解基因相互作用的社团,同时给出每个社团内部和社团之间基因的相互作用情况.

2　基因表达时延的估计

在生物体内,基因表达之间的延时调控现象是客观存在的,如果忽略此时延而进行建模,则无法反映基因之间真实的相互作用过程.由于基因表达数据可视为一系列离散化的时间序列,为更好地理解基因间的调控关系,利用时间序列谱分析方法,可以从频域角度反映基因表达时间序列周期波动特征的全部信息.

假设两个基因表达时间序列分别为 $x(t)$、$y(t)$,待估计的时间延迟为 T,那么互谱的计算公式为:

$$P_{xy}(e^{jw}) = \frac{1}{N} X^*(e^{jw}) Y(e^{jw}) \tag{1}$$

式中,$X^*(e^{jw})$是 $x(t)$的自功率谱密度的共轭,$Y(e^{jw})$是 $y(t)$的自功率谱密度,N 是进行谱分析的点数.根据谱分析理论,二元平稳序列的互谱密度函数反映了两个信号在频域的相互关系,表示它们在每个频率 f 处有多少相同的功率,可以表示它们波形之间的相似性和在时间上的延迟关系.相干系数是两个序列中频率分量为 f 的分量的振幅乘积的标准化均值,定义为:

$$C_{xy}^2(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f) P_{yy}(f)} \tag{2}$$

C_{xy}的取值区间为[0,1],反映了两个序列在频率 f 处的相关程度,相干系数越接近 1,则两序列在频率 f 处越相关.

信号 $x(t)$和 $y(t)$之间的互谱密度函数是复数,它的幅值即幅度谱,是输出和输入之间的平均幅值变化,它的幅角 Φ(相位谱)是两个序列中对应频率分量相位变化的均值,表示一个序列对另一个序列频率成分的主导程度,反映了序列中的各频率分量的相位差,即超前、滞后关系,通常限定区间为[$-\pi,\pi$].在具有高相干性的频率处的相位谱具有很高的可靠性,其值意味着一个基因控制另外一个基因,并且具有一定的时间延迟.设时间延迟为 T(s),相位为 Φ(rad),频率为 f(Hz),则它们之间的关系为:

$$T = \frac{\Phi}{2 * \pi * f} \tag{3}$$

3　基因调控网络的社团划分

3.1　社团结构的定义

如果将复杂系统内部的各个元素抽象为节点,元素之间的关系视为连接,那么就构成了一个复杂网络来反映系统的复杂拓扑结构.随着对网络性质的物理意义和数学特性的深入研究,人们发现许多实际网络都有一个共同性质,即社团结构.每个社团内部的节点之间的连接相对非常紧密,但是各个社团之间的连接却相对来说比较稀疏,这即是社团结构的主要特征[5].对复杂网络进行社团划分有助于分析复杂网络的拓扑结构、理解复杂网络的功能、发现复杂网络中隐藏的规律以及预测复杂网络的行为.

3.2　基因调控网络的社团划分

将各个基因抽象为节点,基因之间的相互关系视为连接,就构成了一个基因相互作用网络,即基因调控网络.如果把所有的基因都投入同一个网络中进行建模,则建立的模型规模将会过大,很可能得到很多接近于0的网络调控系数值,造成计算浪费.而通常情况下只有少数几个基因起调控作用,这几个主导基因各自组成功能团,调控着其它基因的表达,这一现象符合复杂网络中的社团特性.基于以上考虑,拟利用复杂网络的分析思想和方法,将一个复杂的基因调控网络分解成多个社团,每个社团内相互作用比较集中紧密,由极少数的基因调控其它几个基因,而社团间的联系非常稀疏,并且可能出现没有连接的情况.

谱分析法和局部搜索法是两类主要的复杂网络社团划分方法,其中谱分析法具有严密的理论基础,操作简单,被广泛地应用于图形分割等领域.但是,传统的谱分析法不能判断社团的数目,已有研究针对此问题进行了改进,提出了基于 Laplace 图谱和 K 均值聚类的多社团发现方法.该方法首先将 Laplace 矩阵的次小特征值和第三小特征值对应的特征向量投射到二维空间,空间中的每个点对应网络各个节点,然后根据节点聚集情况来确定社团的个数和聚类中心,应用 K 均值聚类算法一次完成多社团的划分[8].本文在此基础上进行了改进,采用规范化 Laplace 矩阵 $\boldsymbol{M}$ 进行操作,具体步骤如下.

假设进行分析的基因集合为$(g_1, g_2, \cdots, g_n)$,n 为基因数量,其对称的邻接矩阵 $\boldsymbol{A} = (a_{ij})_{n*n}$,如果 g_i 与 g_j 两个基因有相互作用,则 $a_{ij} = 1$,否则 $a_{ij} = 0$.网络的 Laplace 矩阵为:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \tag{4}$$

式中,$\boldsymbol{D}$ 是一个对角矩阵,其对角线上的元素 $D_{ij} = \sum_{j=1}^{n} a_{ij}$对应各个节点的度,$n$ 是网络中节点的数目,

即基因的个数.得到规范化 Laplace 矩阵如下:

$$\boldsymbol{M} = \boldsymbol{D}^{-1/2} * \boldsymbol{L} * \boldsymbol{D}^{-1/2} \tag{5}$$

$\boldsymbol{M}$ 的最小特征值为 0,求出 $\boldsymbol{M}$ 的次小特征值和第三小特征值对应的特征向量,将各个节点投影到它们所确定的二维空间,根据节点聚集情况确定社团的数目 k,然后选择 $k-1$ 个最接近于 0 的特征值.在这 $k-1$ 个特征向量中,同一个社团内节点对应的特征向量元素非常接近,将它们进行 K 均值聚类,即可得到最终的社团划分情况.

4 实验结果与分析

4.1 数据选择与预处理

选择酵母细胞周期调控基因中 104 个已知功能有关的基因在 alpha 因子作用下 18 个时间点的表达数据(http://cellcycle-www.stanford.edu/),去除其中缺失数据超过 50%的 3 个基因,选择 101 个基因进行分析.DNA 芯片数据采用基因表达的 Ratio 值,选择采用基因表达的量,即 mRNA 的浓度.因此,首先,将原始实验数据取以 2 为底的指数,以还原基因表达的量,消除负值;然后,用三次样条插值法补齐缺失样本点.酵母细胞周期过程与一般的细胞周期一样,分为 G1、S、G2 和 M 共 4 个阶段.这些基因在不同的细胞周期阶段表达模式不同,对应地也分为不同的时间段.Spellman 等人把这些基因分为 G1、S、S/G2、G2/M 和 M/Gl 共 5 相,对应分为 5 类[9].

4.2 基因表达时延的估计

实验给出了基因 hta2 和 htb1 表达时延的求取过程,以此为例来说明采用谱分析法求时延的有效性.图 1给出了基因 hta2 和 htb1 的原始表达曲线,可以看出两个曲线形状相似,htb1 比 hta2 超前表达.图 2 给出了它们之间的相干谱和相位谱,由图可得,这两个基因在频率 0.25Hz 处具有高相干性.由式(3)可得,hta2 比 htb1 延迟表达 0.7664 分钟.将 htb1 的表达曲线向后平移 0.7664 分钟,如图 3 所示,两条表达曲线在各个转折点处对应的时间点基本吻合.

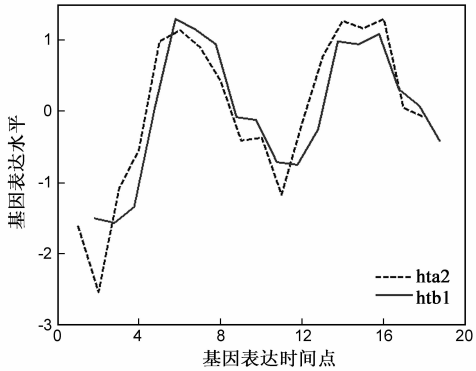


图1 基因hta2和htb1的原始表达曲线

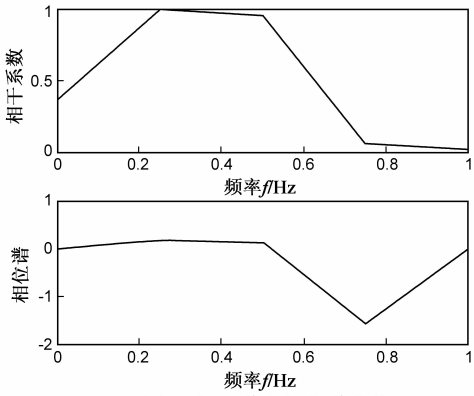


图2 相干谱和相位谱曲线

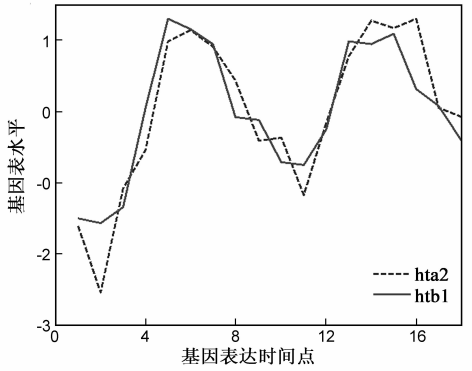


图3 基因hta2和htb1调整后的表达曲线

4.3 基因相互作用邻接矩阵的构建与社团划分

为进行社团划分,首先构造基因相互作用邻接矩阵 $\boldsymbol{A}$.相关系数及其阈值对邻接矩阵的构建以及社团的划分有直接影响,因此应对其进行合理选择.

Pearson 相关系数是一种线性相关系数,该方法将所有的基因表达数据进行分析计算,它反映了基因向量所表示的曲线变化的密切程度[10].Pearson 相关系数的取值范围为[-1,1],为正表示正相关,为负表示负相关.Spearman 秩相关系数是一种非线性相关系数,它并不要求基因表达曲线完全重合或完全相反,只要有相同趋势即可,它比线性相关系数要求更低.信息熵相关系数是一种广义的相关系数,它用于描述两个基因向量有多少共同的信息,它的性质类似于非线性相关系数,但更具有不确定性,即当用不同的方法离散连续变量时就会得到不同的结果,这将对操作带来干扰.综合考虑各种方法的优缺点,选择 Pearson 相关系数来表示基因之间的相关性.

首先根据基因表达时延的估计来调整各个时间序列,求得各个基因之间的 Pearson 相关系数,然后根据相关系数分布情况确定阈值,高于此阈值则认为两个基因之间有相互作用,有 $a_{ij}=1$,否则 $a_{ij}=0$.因此,$\boldsymbol{A}$ 对角线上的元素全部为 0.阈值的大小对于社团的划分影响很大,若阈值较低,则可能会定义一些不相关的基因节

点连接,从而导致划分的社团有些混乱.若阈值过高,就会出现度值为0的节点,从而导致无法再对规范化的Laplace矩阵 $\boldsymbol{M}$ 进行操作.为解决这个问题,本文适当提高阈值,将所有度值为0的节点挑选出来,令其度值为1,然后进行社团划分,这样就能适当提高相关系数阈值,以保证高于此阈值的基因对之间具有相互作用的可靠性,本文设定该阈值为0.65以上.

接下来求出 $\boldsymbol{M}$ 的次小特征值和第三小特征值对应的特征向量,将各个节点投影到它们所确定的二维空间,根据节点聚集情况确定社团的数目.如图4所示,可以看出全部节点可分为5个社团,与Spellman等人的实验结果相符合,将这些节点进行模糊C均值聚类,得到各个社团基因的分布情况.

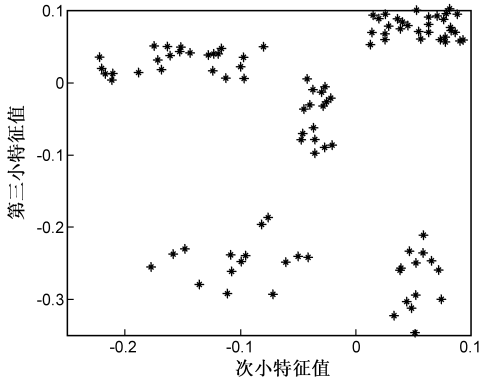


图4　Laplace法划分社团示意图

根据Spellman等人的结论,这5相对应的基因数目分别为53、8、9、15和19,各相基因表达曲线如图5所示.Spellman等人称其测出了95个基因的表达规律,其结果相对于传统实验方法具有91%的正确率[9].从图5可以看出,第2相最混乱,第3相最工整.

采用模糊C均值聚类法对基因进行聚类,如图6所示,每相基因的数目分别为28、27、10、18和18,可以看出每一相基因的总体表达趋势比较整齐.但如前文所分析,仅根据基因表达曲线对其进行聚类,容易忽视生物系统多因素和非线性的特点,不能仅通过聚类对基因进行分组分析.

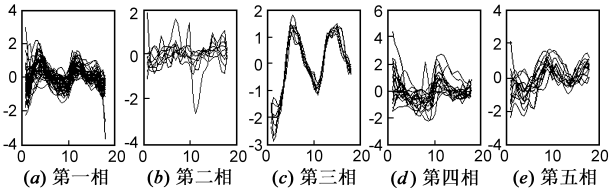


图5　Spellman等人的实验结果

为说明求取基因表达时延对社团划分的积极作用,首先不经过求基因表达时延直接进行社团划分,结果如图7所示,可以看出社团划分比较混乱.增加基因表达时延估计这一操作后的实验结果如图8所示,每相基因的数目分别为32、24、14、16和15,每相整体走势与直接采用模糊 C 均值聚类法得到的结果大体相同,说明采用社团划分法得到的每相基因序列谱相似,符合构建基因调控网络的前提.每相基因之间的调控关系可以从邻接矩阵中获得,从而得到基因调控的细节信息.为方便起见,本文简化了每个社团,给出其连接的示意图,如图9所示.简单的聚类分析只能给出基因的聚类情况,不能提供进一步的调控信息.

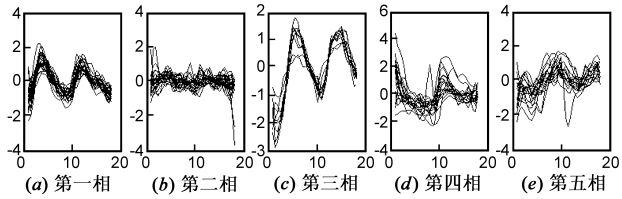


图6　模糊C均值聚类实验结果

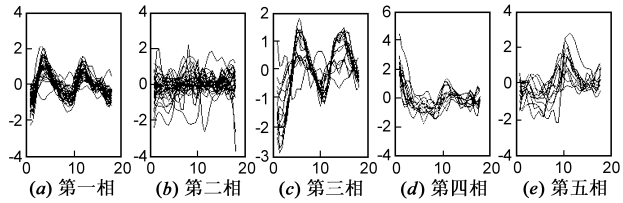


图7　未求时延进行社团划分的实验结果

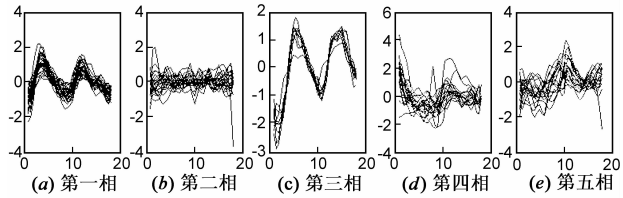


图8　求时延后进行社团划分的实验结果

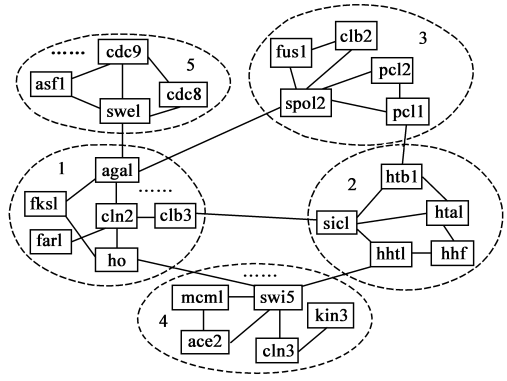


图9　5个基因社团示意图

5　结论

基因调控网络的研究从基因之间相互作用的角度揭示复杂的生命现象,是功能基因组学研究的重要内容,也是当前生物信息学研究的前沿.常用的聚类方法在构建基因调控网络时,只能找到共同调控的基因,不能准确地反映基因之间的相互作用过程和提供基因调控模型的细节.由于复杂网络是分析复杂系统有力的

数学工具,是对复杂系统非常一般的抽象和描述方式,它突出强调了系统结构的拓扑特征.为此,利用复杂网络中的社团结构特性,提出一种基于复杂网络的时延基因调控网络构建方法.首先,采用谱分析法求解各个基因之间的表达时延;其次,基于基因间的相关性分析建立基因相互作用邻接矩阵;然后,进行复杂网络的社团划分,求解基因相互作用的社团,进而根据之前建立的邻接矩阵得到各个社团内部和社团间基因相互作用的细节信息.酵母细胞周期调控基因仿真实验结果表明,与 Spellman 等人和模糊 C 均值聚类方法对比,本文提出的方法不仅能够有效划分基因相互作用的聚类,并且能够进一步提供基因调控模型的细节.

参考文献:

[1] 李辉,王金莲.基于基因表达谱的肿瘤预测模型研究[J].电子学报,2008,36(5):989 – 992.
Li Hui, Wang Jin-lian. Study of tumor molecular prediction model based on gene expression profiles[J]. Acta Electronic Sinica, 2008, 36(5): 989 – 992. (in Chinese)

[2] 张宏怡,张军英.延迟基因调控网络重构问题研究[J].西安电子科技大学学报(自然科学版),2007,34(5):809 – 813.
Zhang Hong-yi, Zhang Jun-ying. Research on the construction of delay gene regulatory network[J]. Journal of Xidian University (Natural Science), 2007, 34(5): 809 – 813. (in Chinese)

[3] L D Han, J Zhu. Using matrix of thresholding partial correlation coefficients to infer regulatory network[J]. Biosystems, 2008, 91(1): 158 – 165.

[4] W P Lee, K C Yang. A clustering-based approach for inferring recurrent neural network as gene regulatory networks[J]. Neuromputing, 2008, 71(4-6): 600 – 610.

[5] 李兵,王浩,李增扬,等.基于复杂网络的软件复杂性度量研究[J].电子学报,2006,34(12A):2371 – 2375.
Li Bing, Wang Hao, Li Zeng-yang, et al. Software complexity metrics based on complex networks[J]. Acta Electronic Sinica, 2006, 34(12A): 2371 – 2375. (in Chinese)

[6] F Comellas, J D Lopez. Spectral reconstruction of complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(25): 6436 – 6442.

[7] J Wu, Y J Tan, H Z Deng, et al. Relationship between degree-rank function and degree distribution of protein-protein interaction networks[J]. Computational Biology and Chemistry, 2008, 32(1): 1 – 4.

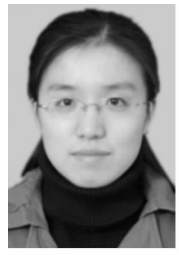
[8] 杨建新,周献中,葛银茂.基于拉普拉斯图谱和 K 均值的多社团发现方法[J].计算机工程,2008,34(12):178 – 180, 183.
Yang Jian-xin, Zhou Xian-zhong, Ge Yin-mao. Method of multi-community finding based on Laplace graph spectrum and K-means[J]. Computer Engineering, 2008, 34(12): 178 – 180, 183. (in Chinese)

[9] P T Spellman, G Sherlock, M Q Zhang, et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization[J]. Molecular Biology of the Cell, 1998, (9): 3273 – 3294.

[10] 易东,杨梦苏,李辉智,等.相关分析在建立基因调控网络中的应用[J].中国卫生统计,2003,20(3):144 – 146.
Yi Dong, Yang Meng-su, Li Hui-zhi, et al. The construction of gene network by correlation analysis[J]. Chinese Journal of Health Statistics, 2003, 20(3): 144 – 146. (in Chinese)

作者简介:

王雪松 女,1974 年生于安徽泗县,2002 年获中国矿业大学控制理论与控制工程专业博士学位,现为中国矿业大学信息与电气工程学院教授,博士生导师.主要研究方向为机器学习、生物信息学等.
E-mail: wangxuesongcumt@163.com



谷阳阳 女,1985 年生于山东济宁,2007 年获中国矿业大学学士学位,现为中国矿业大学控制理论与控制工程专业硕士研究生,研究方向为生物信息学.
E-mail: cumtgyy@126.com

程玉虎 男,1973 年生于安徽淮南,2005 年获中国科学院自动化研究所控制理论与控制工程专业博士学位,现为中国矿业大学信息与电气工程学院副教授.主要研究方向为机器学习和智能系统等.
E-mail: chengyuhu@163.com