高维空间点覆盖方法在物种计算机自动分类中的应用

安 冬[1],王 库[1],王守觉[2]
(1. 中国农业大学,北京 100083; 2. 中国科学院半导体研究所,北京 100083)

摘 要: 长期以来物种的分类主要依靠形态学方法,难以形成计算机自动分类.本文提出了一种基于高维空间几何分析的序列对比方法,并应用该方法对 9个嗜肝病毒科病毒和 14个花椰菜花叶病毒科病毒做出了进化树,结果完全符合国际病毒学命名委员会公布的病毒分类标准.在此基础上,本文探索性的提出了一种基于仿生模式识别的物种自动分类方法,并应用该方法对嗜肝病毒科病毒和花椰菜花叶病毒科病毒做了自动分类,正确分类率分别达到 100%和 94%.

关键词: 高维空间几何分析; 序列对比; 仿生模式识别; 物种自动分类
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2006) 02-0277-05

A Alignment-Free Sequence Comparison Method Based on Whole Genomes and Its Application to Virus Phylogeny

AN Dong[1], WANG Ku[1], WANG Shou-jue[2]
(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China;
2. Institute of Semiconductors, CAS, Beijing 100083, China)

Abstract: For a long time, the classification of species depends mainly on morphologic methods. And it is difficult to realize automatic species classification by using computer. This article puts forward a alignment-free sequence comparison method based on whole genomes and draws phylogeny trees for 9 Hepadnaviridae viruses and 14 Caulimoviridae viruses with this method. The results match the standard of virus classification published by ICNV. On this foundation, the article exploringly puts forward an automatic classification method of species based on pattern recognition and uses this method to do the automatic classification for Hepadnaviridae viruses and Caulimoviridae viruses. The correct classification rate reaches 100% and 94% respectively.

Key words: sequence comparison; automatic classification of species; virus phylogeny; whole genomes

1 引言

近十几年来,生物科学得到迅猛发展,生物学相关数据的积累速度大大超出了人们的想象.面对生物信息的爆炸性增长,现有的信息收集、存储、分析和处理方法已远远不能满足实际工作的需要,亟待改进和更新.

1859年达尔文在《物种起源》中提出的进化学说是对人类自然科学和自然哲学发展的重大贡献.长期以来物种的分类主要依靠形态学方法,难以形成计算机自动分类.

随着分子生物学和生物信息学的发展,从分子水平上研究物种进化成为可能.早期的研究工作主要是利用不同物种中同一基因序列的异同或其编码的氨基酸序列的异同来研究生物的进化,以及通过对比不同物种中同一蛋白质的结构来研究生物的进化.以上的研究已经积累了大量工作的经验[1,2].

通常从分子水平上研究物种进化是通过比较不同物种之间的相关氨基酸序列来进行的,然而氨基酸的编码具有简约性,即编码同一氨基酸的三位核苷酸可以不同,这种简约可能伴随着有用信息的丢失,因此许多直接与进化相关的信息往往通过核苷酸反映出来.所以有人说:"No other biological sequence can bring more phylogenetic information than genome".近年来,由于较多生物完整基因组测

收稿日期: 2005-07-29;修回日期: 2005-10-25

序任务的完成,从完整基因组角度研究物种进化成为可能.

从完整基因组角度研究物种进化时一个关键的问题是怎样进行不同物种完整基因组之间的序列比较.早期工作中常采用的序列对位排列方法(sequence alignment)并不适用于不同物种完整基因组之间的序列比较[3,4].现今急需快速、有效的针对完整基因组的序列比较方法.本文提出了一种基于高维空间几何分析的序列对比方法,可以从不同物种的完整基因组序列中提取出长度相同的16维特征向量来表征这个物种,进而可对这些特征向量进行对比分析.为了验证该方法的有效性,本文基于该方法对9个嗜肝病毒科(Hepadnaviridae)病毒和14个花椰菜花叶病毒科(Caulimoviridae)病毒做出了进化树,结果完全符合国际病毒学命名委员会(ICNV)公布的病毒分类标准.

人们常用进化树来表征物种的分类及进化关系,发展了很多的建树方法.但进化树模型有其一定的弊端,例如加入一个新的物种就必须完全重新计算等.本文在基于高维空间几何分析的序列对比方法基础上采用PCA(principal components analysis,主元分析)对特征向量进行降维,探索性的使用一种两维图模型来表征物种的分类,并应用该方法对9个嗜肝病毒科病毒、14个花椰菜花叶病毒科病毒、及40个逆转录病毒科(Retroviridae)病毒做了分类,取得了一定的效果.

为了进一步研究物种的计算机自动分类,在基于高维空间几何分析的序列对比方法基础上,本文提出了一种基于仿生模式识别的物种自动分类方法,并应用该方法对嗜肝病毒科病毒和花椰菜花叶病毒科病毒做了自动分类,正确分类率分别达到100%和94%.

2 材料和方法

(1)材料

本文所采用的病毒完整基因组为GenBank数据库中嗜肝DNA病毒科的9种病毒、花椰菜花叶病毒科的14种病毒和逆转录病毒科的6种病毒的完整基因组核苷酸序列,嗜肝DNA病毒科包括3种禽嗜肝DNA病毒(Avihepadnavirus)和6种正嗜肝DNA病毒(Orthohepadnavirus);花椰菜花叶病毒科包括7种花椰菜花叶病毒(Caulimovirus)和7种病毒杆状DNA病毒(Badnavirus).

(2)基于高维空间几何分析的序列对比方法

(a)将一个完整基因组序列,分解为同样长度的四个子序列,每个子序列(SA, SC, SG, ST)分别表示对应字符(A, C, G, T)在原序列中的出现情况.如一个为AGCTAGCT的序列,可以被分解为4个子序列:SA、SC、SG、ST.

SA: 10001000

SC: 00100010

SG: 01000100

ST: 00010001

(b)各子序列中的1代表在原序列中该位置上有某一特定字符(A或C或G或T)出现,0代表在原序列中该位置上没有某一特定字符(A或C或G或T)出现.

(c)若原序列长度为m(即原序列共有m个字符),那么被分解出来的4个子序列(SA, SC, SG, ST)可以看成4个m维向量,即m维空间中的4个点,计算这4个向量之间的线性相关系数和这4个点之间的欧氏距离,共得到12个数据,即6个距离数据、6个夹角数据.

(d)计算此完整基因组序列中A、C、G、T四个字母出现的频率,得到4个数据.

(e)将第c步和第d步计算所得的共16个数据组成一个16维的新向量,作为表征此完整基因组序列的特征向量(即16维特征空间中的一个特征点).

(f)重复步骤a至e,共计算出表征n个完整基因组序列的n个特征向量,即在16维特征空间里得到了n个特征点,计算任两点之间的欧氏距离,得到一个n×n的距离矩阵.

(g)用PHYLIP 3.6软件包中的NEIGHBOR软件来构建对应此距离矩阵的种系进化树.为了评价树的拓扑结构的鲁棒性,采用Jack-knife算法随机产生100个输入数据集,并用PHYLIP3.6软件包中的CONSENCE软件来构建共有树.

(3)两维图模型

(a)据(2)小节f步,得到表征n个完整基因组序列的n个特征向量($V_1, V_2, \cdots, V_n$)将这些特征向量组成一个16×n的矩阵H.

$$H = [V_1, V_2, \cdots, V_n]$$

$$V_i = [v_{i1}, v_{i2}, \cdots, v_{i16}]^T,\ i = 1\cdots n \tag{1}$$

其中,v_j ($j = 1\cdots16$)为(2)小节c步和d步计算得到的16个数据.

(b)对矩阵H进行PCA处理.PCA的目的是在16维特征空间中找到一组正交向量(m个),这组向量可最大可能的表示出数据的方差.将数据从原来的n维空间投影到这组正交向量所组成的m维子空间上,从而完成维数压缩的作用.具体算法如下[10]

计算

$$\Sigma = \sum_{i=1}^{n} (V_i - \mu) \tag{2}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} V_i \tag{3}$$

对矩阵Σ进行SVD分解(Singular Value Decomposition)

$$\Sigma = U\Lambda^{1/2}V \tag{4}$$

得到

$$U = [u_1, u_2, \cdots, u_n] \in R^{16*n} \tag{5}$$

$$U^T U = I \tag{6}$$

$$\Lambda = \mathrm{diag}[\lambda_1, \lambda_2, \cdots, \lambda_n] \in R^{n*n} \tag{7}$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \tag{8}$$

$$V = [v_1, v_2, \cdots, v_n] \in R^{n*n} \quad (9)$$

$$VV^T = I \quad (10)$$

根据 λ值的大小选取前 2个特征值所对应的 2个正交归一的特征向量 U

$$U = [u_1, u_2] \in R^{16*2}$$

(c)计算 $G = U^T \Sigma$,将原 16维特征向量压缩为 2维新特征向量

$$G = [VN_1, VN_2, \cdots, VN_n]$$

$$VN_i = [vn_{i1}, vn_{i2}]^T \quad i = 1 \cdots n \quad (11)$$

其中 VN_i ($i = 1 \cdots n$)为表征 n个完整基因组序列信息的 n个新特征向量.

(d)以 VN_{i1}为横坐标, VN_{i2}为纵坐标,则 VN_i ($i = 1 \cdots n$)即为平面上的 n个点.

(4)基于仿生模式识别的物种自动分类方法

(a)使用 (2)小节 e步中计算得到的 16维向量作为表征完整基因组序列的特征向量,即 16维特征空间中的一个样本点. 选取 4个嗜肝 DNA 病毒科和 8个花椰菜花叶病毒科病毒的完整基因组序列经特征提取得到特征向量,作为训练样本集. 选取 5个嗜肝 DNA 病毒科和 6个花椰菜花叶病毒科病毒的完整基因组序列经特征提取得到特征向量,作为第一测试样本集. 选取 6个逆转录病毒科病毒的完整基因组序列经特征提取得到特征向量,作为第二测试样本集.

(b)神经网络的构造和训练:采用高维空间点覆盖的方法构造、训练 2个网络,分别用于识别嗜肝 DNA 病毒科病毒和花椰菜花叶病毒科病毒.

从高维空间几何分析的角度来看,一个神经元可以构造出一个复杂的封闭几何形体,多个神经元组合起来的人工神经网络可以实现高维空间复杂几何形体的近似覆盖. 下面我们就采用一个 2权值神经元作为基本覆盖单元,用多个 2权值神经元组合起来实现两类病毒的神经网络覆盖区.

多权值神经元可以表示为:

$$Y = f[\Phi(X, W_1, W_2, \cdots, W_m) - Th] \quad (12)$$

其中 X为输入矢量, $W_1, W_2, \cdots, W_m$ 为权值矢量, Φ 表示输入矢量 X与权值矢量 $W_1, W_2, \cdots, W_m$ 之间的函数关系, Th为阈值, f为判别函数.

当 m = 2时,上式为一 2权值神经元,我们将其命名为 HSN (超香肠形神经元),其表达式为,

$$Y = f[\Phi(X, W_1, W_2) - Th]$$

$$\Phi(X, W_1, W_2) = \| X - \theta_{(W_1, W_2)} \| \quad (13)$$

其中 $\theta_{(W_1, W_2)}$表示由 n维空间中两点 W_1、W_2 确定的有限一维线段,

$$\theta_{(W_1, W_2)} = \{ Y | Y = \alpha W_1 + (1 - \alpha) W_2, \alpha \in [0, 1] \} \quad (14)$$

$\Phi(X, W_1, W_2)$表示 n维空间中的点 X到 n维空间中的有限一维线段 $\theta_{(W_1, W_2)}$的欧式距离.

判别函数 f为:

$$f(x) = \begin{cases} 1, & \text{当 } x \le 0 \\ -1, & \text{当 } x > 0 \end{cases} \quad (15)$$

HSN 神经元的覆盖区域实际上是 n维空间中一个有限一维线段和超球的拓扑乘积,超球的半径为阈值 Th.

构造各类高维空间点覆盖区的具体步骤如下:

步骤 1　设某类病毒所有的构网样本点集合为 $\alpha = \{A_1, A_2, \cdots, A_N\}$, N为样本点总数.

在 16维特征空间中计算所有点两两之间的欧式距离,找出距离最小的两个点,记为 B_{11}、B_{12}. 这样在 16维特征空间中就由点 B_{11}、B_{12}构成第一个一维线段$\overline{B_{11}B_{12}}$,记作 θ_1. 用一个 HSN 神经元来覆盖这个线段,其覆盖范围为:

$$P_1 = \{ X | \rho_{X\theta_1} \le Th, X \in R^n \}$$

$$\theta = \{ Y | Y = \alpha B_{11} + (1 - \alpha) B_{12}, \alpha \in [0, 1] \} \quad (16)$$

其中 $\rho_{X\theta_1}$表示点 X到空间 θ_1 的距离.

步骤 2　对于已构造好的几何形体 P_1,判断剩余各点是否被 P_1 覆盖. 若在 P_1 覆盖范围内,则排除该点;对于在 P_1 覆盖范围外的各点,按照步骤 1的方法,找出离 B_{12}距离最近的一点,记作 B_{13},这样 B_{12}与 B_{13}就构成第二个线段$\overline{B_{11}B_{12}}$,记作 θ_2. 同样,用一个 HSN 神经元来覆盖这个线段,其覆盖范围为:

$$P_2 = \{ X | \rho_{X\theta_2} \le Th, X \in R^n \}$$

$$\theta = \{ Y | Y = \alpha B_{12} + (1 - \alpha) B_{13}, \alpha \in [0, 1] \} \quad (17)$$

其中 $\rho_{X\theta_2}$表示点 X到空间 θ_2 的距离.

…

步骤 i　在剩余点中排除包含在前面共 (i - 1)个 HSN 神经元覆盖范围内的样本点,在覆盖范围外的样本点中,找出离第 $B_{1(i-1)}$点距离最近的点,记作 B_{1i},这样 $B_{1(i-1)}$与 B_{1i}就构成第 i个线段$\overline{B_{1(i-1)}B_{1i}}$,记作 θ_i,同样,用一个 HSN 神经元来覆盖这个线段,其覆盖范围为:

$$P_i = \{ X | \rho_{X\theta_i} \le Th, X \in R^n \}$$

$$\theta_i = \{ Y | Y = \alpha B_{1(i-1)} + (1 - \alpha) B_{1i}, \alpha \in [0, 1] \} \quad (18)$$

…

直到处理完所有的构网样本点.

最终共产生 m个 HSN 神经元,每一类病毒的覆盖区域就是这些神经元覆盖区域的并集:

$$P = \bigcup_{i=1}^{m} P_i \quad (19)$$

将距离待识样本点 X最近的那类单词音节高维空间覆盖区所属类别,作为待识样本点 X的所属单词音节类别.

(c)样本识别:用两个构造好的网络识别第一、第二测试样本集中的所有样本. 待识别样本落入哪个网络覆盖区,识别结果即为哪个网络所代表的类别. 如若待识别样本没有落入任何网络覆盖区,则该样本不属于任何网络所代表的类别.

判别某一待识别样本点是否属于某单词音节高维空间覆盖区的方法:

计算待识别样本点 X到各类病毒高维空间覆盖区中各 HSN 神经元覆盖区域的距离:

$$\rho = \| X - \theta_{(W_i, W_j)} \| \quad (20)$$

则待识别样本点到第类病毒高维空间覆盖区的距离为

$$\rho_i = \min_{j=1}^{M_i} \rho_{ij} \quad i = 1, \cdots, 11 \quad (21)$$

其中 M_i 为构成第 i类病毒高维空间覆盖区的 HSN 神经元个数,ρ_{ij}为待识别样本点 X到第 i类病毒高维空间覆盖区中第 j个 HSN 神经元覆盖区域的距离.

3 结果和讨论

(1)结果

本文应用基于高维空间几何分析的序列对比方法对 9个嗜肝病毒科 (Hepadnaviridae)病毒和 14个花椰菜花叶病毒科 (Caulimoviridae)病毒做出了进化树,结果完全符合国际病毒学命名委员会 (ICNV)公布的病毒分类标准. 结果见图 1、和图 2.

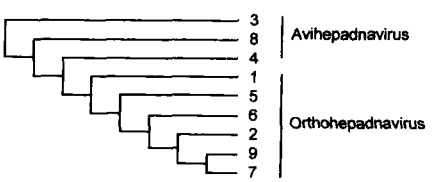

图 1 使用基于高维空间几何分析的序列对比方法做出的 3个禽嗜肝 DNA 病毒属 (Avihepadnavirus)病毒和 6个正嗜肝 DNA 病毒属 (Orthohepadnavirus)病毒的进化树

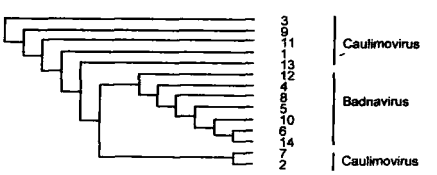

图 2 使用基于高维空间几何分析的序列对比方法做出的 7个花椰菜花叶病毒属 (Caulimovirus)病毒和 7个杆状 DNA 病毒属 (Badnavirus)病毒的进化树

本文应用两维图模型对嗜肝病毒科、花椰菜花叶病毒科及逆转录病毒科做了自动分类. 结果见图 3、4、5、6.

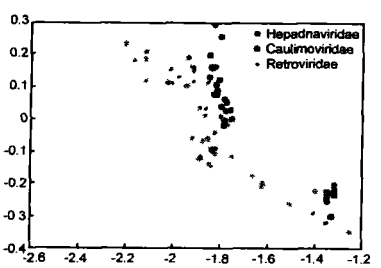

图 3 使用两维图模型做出的嗜肝病毒科、花椰菜花叶病毒科及逆转录病毒科病毒种的分类

本文使用基于仿生模式识别的物种自动分类方法对嗜肝病毒科病毒和花椰菜花叶病毒科病毒做了自动分类,实验结果如下表. 其中用来识别嗜肝 DNA 病毒科病毒种的 HSN 网络对嗜肝病毒科病毒和花椰菜花叶病毒科病毒的正确识别率为 100%,对逆转录病毒科病毒的正确据识率为 100%;用来识别花椰菜花叶病毒科病毒种的 HSN 网络对嗜肝病毒科病毒和花椰菜花叶病毒科病毒的正确识别率为 96%,错误据识率为 4%,对逆转录病毒科病毒的正确据识率为 96%,错误识别率为 6%.

表 1 基于仿生模式识别的物种自动分类方法结果

	正确识别率	正确据识率	错误识别率	错误据识率
识别嗜肝 DNA 病毒科病毒种的 HSN 网络	100%	100%	0%	0%
识别花椰菜花叶病毒科病毒种的 HSN 网络	94%	94%	6%	6%

我们使用同样的训练样本集采用 RBF核的支撑向量机进行学习,并使用同样的测试样本集进行测试,结果表明对嗜肝病毒科病毒和花椰菜花叶病毒科病毒的正确识别率为 65.52%,对逆转录病毒科病毒的正确据识率为 0%,即所有未经训练的类别均会被错误的分类,而不会正确据识. 这证明了基于仿生模式识别的物种自动分类方法在小样本集情况下的绝对优越性.

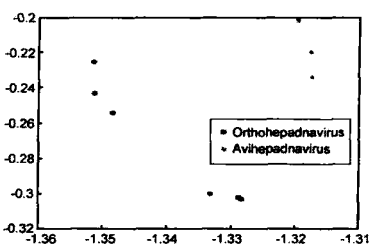

图 4 使用两维图模型做出的 3个禽嗜肝 DNA 病毒属 (Avihepadnavirus)病毒和 6个正嗜肝 DNA 病毒属 (Orthohepadnavirus)病毒的分类

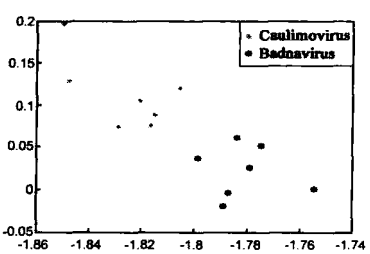

图 5 使用两维图模型做出的 7个花椰菜花叶病毒属 (Caulimovirus)病毒和 7个杆状 DNA 病毒属 (Badnavirus)病毒的分类

(2)讨论

(a)怎样对病毒进行分类是生物学上一个长期存在的问题,这是因为病毒可以利用的形态学特征十分有限,而在分子水平上研究病毒的分类更为困难,因为在病毒的基因组里不存在研究其他生物分类时常用的 16s rRNA 基因. 近年来,由于较多病毒完整基因组测序任务的完成,从完整基因组角度研究病毒

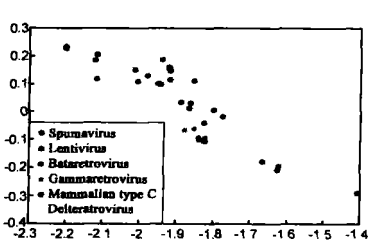

图 6 使用两维图模型做出的逆转录病毒科病毒分类

进化成为可能.从完整基因组角度研究病毒进化时一个关键的问题是怎样进行不同病毒完整基因组之间的序列比较.早期工作中常采用的序列对位排列方法并不适用于不同病毒完整基因组之间的序列比较.近年来,人们发展了很多的不基于序列对位排列的序列对比方法.这些方法主要分为两大类:第一类方法是基于字符或字符串出现频率的,第二类方法是基于 Kolmogorov 复杂性和 Chaos 理论的[5].本文从几何的角度出发思考这个问题,提出了一种基于高维空间几何分析的序列对比方法,这种方法概念清晰、计算简便,是一种快速、有效的针对完整基因组的序列比较方法.

(b)仿生模式识别是王守觉院士提出的一种模式识别新理论[6~9],完全不同于传统的模式识别理论.传统模式识别把不同类样本在特征空间中的最佳划分作为目标,最具代表性的就是支持向量机(SVM)理论;而仿生模式识别则以同一类样本在特征空间中分布的最佳覆盖作为目标.仿生模式识别理论通过分析某类样本点在高维空间中的分布情况,利用高维空间复杂几何形体对其进行覆盖.一个神经元可以是一个的复杂的封闭超曲面,多个神经元组合起来的人工神经网络就可以实现高维空间复杂几何形体覆盖.因而,人工神经网络是实现仿生模式识别的十分合适的手段.在本文中作者将仿生模式识别应用在生物领域,探索性的提出了一种基于仿生模式识别的物种自动分类方法,取得了较好的分类效果.

参考文献:

[1] Wen-Hsiung Li Molecular Evolution[M]. Sinauer Associates. 1997. 3 - 11.

[2] Masatoshi Nei, Sudhir Kumar. Molecular Evolution and Phylogenetic[M]. Oxford University Press, 2000. 77 - 83.

[3] Webb Miller. Comparison of genomic DNA sequences: solved and unsolved problems [J]. Bioinformatics, 2001, 17(5): 391 - 397.

[4] Bailin Hao, Ji Qi. Prokaryote phylogeny without sequence alignment: From avoidance signature to composition distance[A]. Proceedings of the Computational Systems Bioinformatics [C]. Los Alamitos: IEEE, 2003. 25 - 56.

[5] Susana Vinga, Jonas Almeida. Alignment-free sequence comparison——a review [J]. Bioinformatics, 2003, 19(4): 513 - 523.

[6] 王守觉,等. 人工神经网络的多维空间几何分析及其理论 [J]. 电子学报, 2002, 30(1): 1 - 4.
WANG Shou-jue, WANG Bai-nan. Analysis and theory of high-dimension space geometry for artificial neural networks[J]. Acta Electronica Sinica, 2002, 30(1): 1 - 4.

[7] 王守觉. 仿生模式识别(拓扑模式识别)——一种模式识别新模型的理论与应用 [J]. 电子学报, 2002, 30(10): 1417 - 1420.
WANG Shou-jue. Bionic(Topological) Pattern Recognition——A new model of pattern recognition theory and its applications [J]. Acta electronica sinica, 2002, 30(10): 1417 - 1420.

[8] 王守觉,等. 基于仿生模式识别的多镜头人脸身份确认系统研究 [J]. 电子学报, 2003, 31(1): 1 - 3.
WANG Shou-jue, XU Jian, WANG Xian-Bao, QIN Hong. Multi-camera human-face personal identification system based on the biomimetic pattern recognition[J]. Acta Electronica Sinica, 2003, 31(1): 1 - 3.

[9] 王守觉,等. 通用神经网络硬件中神经元基本数学模型的讨论 [J]. 电子学报, 2001, 29(5): 577 - 580.
WANG Shou-jue, LI Zao-zhuo, et al. Discussion on the basic mathematical models of neurons in general purpose neurocomputer[J]. Acta Electronica Sinica, 2001, 29(5): 577 - 580.

[10] 斯华龄(美). 智能视觉图像处理 [M]. 上海:上海科技教育出版社, 2002. 125 - 128.

作者简介:

安　冬　女,1977年生,博士,研究方向为模式识别、人工神经网络、信号处理,生物信息学. E-mail: andong@semi.ac.cn.

王　库　男,教授,博导,中国农业大学电子系主任,研究方向为模式识别、智能化检测与控制技术等.