第 32 卷 第 1 期 华 中 农 业 大 学 学 报 Vol. 32 No. 1
2013 年 1 月 Journal of Huazhong Agricultural University Jan. 2013,125～129

基于 Fisher 判别法的一种 DNA 序列分类方法

杨莉萍[1] 路松峰[2] 胡和平[2] 黄 钰[1]

1. 华中农业大学理学院,武汉 430070; 2. 华中科技大学计算机学院,武汉 430074

摘要 针对 DNA 序列分类的分属问题,提出采用 Fisher 判别法进行分类。根据氨基酸分子的性质,构建 DNA 序列对应的特征向量空间,然后对训练集中的 A、B 两类 DNA 序列提取特征,建立 Fisher 判别函数。应用该判别函数对测试集中的 182 个 DNA 序列进行分类实验,结果表明该方法具有很好的分类准确度。

关键词 生物信息学; Fisher 判别法; DNA 序列分类; 特征提取; 特征向量空间

中图分类号 Q 751 **文献标识码** A **文章编号** 1000-2421(2013)01-0125-05

Fisher 判别法是寻找一个投影把高维空间的样本投影到一条直线上,使得同类的样本集中在一起,不同类的样本尽量分开的一种模式识别方法。投影后数据点就变得比较密集,从而可以克服由于特征空间的维数较高而引起的“维数灾难”。该方法根据类间距离最大、类内距离最小的原则确定判别函数,再依据建立的判别函数判定样本的类别。

DNA 全序列结构的研究是生物信息学的一个重要课题,而 DNA 序列分类是研究 DNA 全序列结构的基础。2000 年 6 月以来,随着人类基因组计划中 DNA 全序列草图的完成,人们对 DNA 序列分类问题作了很多研究,建立了一些数学模型,提出了一些分类方法,例如基于图像[1],距离、人工神经网络模型的方法[2],支持向量机的分类方法[3]等。上述方法有其自身的特点,同时也存在一定的局限性。为此,笔者提出一种基于统计学理论 Fisher 判别法的 DNA 序列分类方法。

1 Fisher 判别法原理

1.1 Fisher 判别的求解方法

把 n 维空间的数据投影到一条直线上,这有可能会使在 n 维空间里分得很开的样本集混杂在一起。Fisher 判别法的目标是要找到某个方向,使在这个方向的直线上,样本的投影能分开得最好。

假设有一集合包含 N 个 n 维的样本 $x_1, x_2, \cdots, x_N$,其中 N_1 个样本属于 ω_1 类,N_2 个样本属于 ω_2 类。若对样本的分量作线性组合,则可得标量 $y=W^T x$。

这样得到集合 $y_1, y_2, \cdots, y_N$。从几何上看,W 的幅度是无实际意义的,重要的是 W 的方向。即希望落在直线上的标以 ω_1 类的样本和标以 ω_2 类的样本的投影能分得很开,而不是混杂在一起。

样本投影之间的分离性可用样本均值之差来度量。设 m_i 是第 i 类 n 维样本的均值:

$$m_i=\frac{1}{N_i}\sum_{j=1}^{N_i}x_j,\ i=1,2 \tag{1}$$

而投影点的样本均值 m_i^* 为

$$m_i^*=\frac{1}{N_i}\sum_{j=1}^{N_i}y_j=W^T m_i,\ i=1,2 \tag{2}$$

为了使类别分离得好,应使同类样本的投影比较密集,这个密集程度可用类内离散度来进行度量。定义同一类样本投影的类内离散度为:

$$S_i^{*2}=\sum_{j=1}^{N_i}(y_j-m_i^*)(y_j-m_i^*)^T,\ i=1,2 \tag{3}$$

$S_1^{*2}+S_2^{*2}$ 称为投影样本的总的类内离散度。所谓 Fisher 线性判别函数被定义为这样的一个线性函数

$$y=W^T x \tag{4}$$

它能使判决函数:

$$J(W)=\frac{|m_1^*-m_2^*|^2}{S_1^{*2}+S_2^{*2}} \tag{5}$$

达到极大。显然,为使 J 最大,应使两类均值之差尽可能大,而各类的类内离散度尽可能小。

定义样本类内离散度矩阵 S_i 和总类内离散度矩阵 S_W 如下:

$$S_i=\sum_{j=1}^{N_i}(x_j-m_i)(x_j-m_i)^T,\ i=1,2 \tag{6}$$

收稿日期: 2012-09-20

基金项目: 国家自然科学基金资助项目(10876012)和中央高校基本科研业务费专项资金(52204-10040)

杨莉萍,讲师,博士研究生. 研究方向:生物信息学、高性能计算. E-mail: teresa@mail.hzau.edu.cn

$$S_W = S_1 + S_2 \tag{7}$$

由于

$$S_i^{*2} = \sum_{j=1}^{N_i} (W^T x_j - W^T m_i)^2 = W^T S_i W, i = 1, 2 \tag{8}$$

因此

$$S_1^{*2} + S_2^{*2} = W^T S_W W \tag{9}$$

相似的

$$|m_1^* - m_2^*|^2 = W^T S_B W \tag{10}$$

其中

$$S_B = (m_1 - m_2)(m_1 - m_2)^T \tag{11}$$

矩阵 S_W 称为总类内离散度矩阵，S_B 称为类间离散度矩阵。引入 S_B 和 S_W 后，将判决函数 J 写成：

$$J(W) = \frac{W^T S_B W}{W^T S_W W} \tag{12}$$

容易看出，使 J 达到极大的向量 W 必须满足：

$$S_W^{-1} S_B W = \lambda W \tag{13}$$

在此由于 $S_B W$ 总是在 $m_1 - m_2$ 方向上的，所以没有必要求出 $S_W^{-1} S_B$ 的特征值和特征向量，可以立即把解写成：

$$W = S_W^{-1}(m_1 - m_2) \tag{14}$$

将(14)式代入(4)式即可得到 Fisher 判别函数[4]。

1.2 Fisher 判别效果检验

为考查以上判别方法是否优良[5]，采用下述的回代估计法计算误判率。设有 m 个总体 G_1，G_2，…，G_m。来自总体 G_i 容量为 n_i 的训练样本 $X_\alpha^{(i)} = (x_{a1}^{(i)}, x_{a2}^{(i)}, \cdots, x_{ap}^{(i)})^T$（其中，$\alpha = 1, 2, \cdots, n_i$；$i = 1, 2, \cdots, m$）。将全部训练样本作为新样本，依次代入已创建的判别函数中，并利用判别准则进行判别，此过程被称为回判。用 n_{ij} 表示将属于总体 G_i 的样本误判为总体 G_j 的个数，设总的误判个数为 N，则误判率 η 的回代估计为：

$$\eta = \frac{N}{n_1 + n_2 + \cdots n_m} \tag{15}$$

2 基于 Fisher 方法的 DNA 序列分类

2.1 DNA 序列的特征提取

DNA(脱氧核糖核酸)是存在于生物细胞的细胞核中携带遗传信息的分子[6]，遗传分子决定了蛋白质的结构，这些分子是生命得以延续的重要物质。DNA 分子是由 2 条多核苷酸链组成的，这 2 条链通过碱基互补配对相互缠绕形成双螺旋结构，其中一条链上的嘌呤总是以氢键与另一条链上的嘧啶相结合，腺嘌呤(A)与胸腺嘧啶互补(T)、鸟嘌呤(G)与胞嘧啶(C)互补。基于这种互补性，可以用一条核苷酸链上的 A、T、C、G 这 4 种碱基形成的线性序列来表示 DNA 序列。

为描述方便，将与 RNA(核糖核酸)上密码子对应的 DNA 上相邻的 3 个碱基称为 DNA 上的密码子。以下将 RNA 中 64 个密码子对应的 20 个氨基酸信息，按一一对应关系，对应到 DNA 的密码子中进行编码(表 1)。如表 1 所示：(1)T，A，C，G 的不同组合构成64种密码子。表1中对每个密码子

表 1 氨基酸与对应的密码子及其编号

Table 1 Amino acid and corresponding codons with numbers

氨基酸 Amino acid	序号 Number	对应的密码子及其编号 Corresponding codons with number
苯丙氨酸 Phe	1	AAA (1) AAG (2)
亮氨酸 Leu	2	AAT (3) AAC (4) GAA (5) GAG (6) GAT (7) GAC (8)
丝氨酸 Ser	3	AGA (9) AGG(10) AGT (11) AGC (12) TCA (13) TCG (14)
酪氨酸 Tyr	4	ATA (15) ATG(16)
半胱氨酸 Cys	5	ACA (17) ACG(18)
色氨酸 Trp	6	ACC (19)
脯氨酸 Pro	7	GGA (20) GGG (21) GGT (22) GGC (23)
组氨酸 His	8	GTA (24) GTG (25)
谷氨酰胺 Gln	9	GTT (26) GTC (27)
精氨酸 Arg	10	GCA (28) GCG (29) GCT (30) GCC (31) TCT (32) TCC (33)
异亮氨酸 Ile	11	TAA (34) TAG(35) TAT (36)
苏氨酸 Thr	12	TGA (37) TGG(38) TGT (39) TGC (40)
天冬酰胺 Asn	13	TTA (41) TTG (42)
赖氨酸 Lys	14	TTT (43) TTC (44)
缬氨酸 Val	15	CAA (45) CAG(46) CAT (47) CAC (48)
丙氨酸 Ala	16	CGA (49) CGG (50) CGT (51) CGC (52)
天冬氨酸 Asp	17	CTA (53) CTG (54)
谷氨酸 Glu	18	CTT (55) CTC (56)
甘氨酸 Gly	19	CCA (57) CCG (58) CCT (59) CCC (60)
甲硫氨酸 Met	20	TAC (61)
终止 Termination	21	ATT (62) ATC (63) ACT (64)

用数字编号，于是可计算出不同密码子在 DNA 序列片段中出现的频率(某种密码子个数/该 DNA 片段中密码子总数)。这样对于每个 DNA 片段，就可得到一个代表其每种密码子出现频率的 64 维向量[7]。(2)64 种密码子构成 20 种氨基酸，和一种编号为 21 的终止码。从氨基酸种类数来分，DNA 序列片段可用一个 21 维向量来表示，向量中每一元素表示每种固定种类氨基酸的含量。

显然，无论上述的 64 维向量或 21 维向量作为分类器的特征向量维数都太高，不易处理。于是根据氨基酸中 R 侧链的酸碱性，将 20 种氨基酸进一步分成 3 类[8]：酸性氨基酸、碱性氨基酸、中性氨基酸。再加上终止信息码，共 4 类(表 2)。由表 2 可知：20 种氨基酸分成 3 类，加上编号为 4 的终止信息。则可以选择每一 DNA 序列中 4 类氨基酸的含量作为分类特征。这样所有的 DNA 序列就构成一个 4 维的特征向量空间。

表 2　氨基酸分类

Table 2　Types of amino acid

氨基酸类型 Types of amino acid	编号 Number	对应氨基酸的序号 Number of amino acid	对应密码子种类的数目 Quantity of codons types
酸性氨基酸 Acidic amino acid	1	17,18	4
碱性氨基酸 Alkaline amino acid	2	8,10,14	10
中性氨基酸 Neutral amino acid	3	1,2,3,4,5,6,7,9,11,12,13,15,16,19,20	47
终止 Termination	4	21	3

2.2　DNA 序列特征向量的产生

对于一个未表明是全序列还是序列片段的 DNA 序列来说，从哪一位碱基开始解读密码子有 3 种不同的方式：从第 1 位开始、第 2 位开始、第 3 位开始解读(从第 4 位开始则与第 1 位重复，依此类推)。例如，对一个 DNA 序列 cggaggacaaac 提取特征，按照 3 种不同的解读方式，结果为 (cgg)(agg)(aca)(aac) 或 (gga)(gga)(caa) 或 (gag)(gac)(aaa)。这 3 种解读方式产生 3 种不同的密码信息，但为了保证生物遗传性状的稳定(子代必须具备与父代相同的 DNA 信息，这主要由碱基的排序来体现)，故在生物体内解读 DNA 的方式应是确定的。因此，以下用概率统计的思想，对 DNA 序列选取一种较为合理的解读方式。

设 $\vec{P}=\left[\frac{4}{64},\frac{10}{64},\frac{47}{64},\frac{3}{64}\right]$，它表示理论上各类氨基酸对应的密码子种类出现的概率。式(16)定义的 D_{ij}，表示该 DNA 序列在第 i 种解读方式下($i=1,2,3$) 第 j 类氨基酸($j=1,\cdots,4$) 在本序列中出现的频率。定义 4 维向量 $\vec{P}_i=(D_{i1},\cdots,D_{i4})$，表示第 i 种解读方式下，4 类氨基酸在该序列中各自出现的频率。故 $E_i=\vec{P}\cdot(\vec{P}_i)^T$ 表示第 i 种解读方式下的数学期望。显然 E_i 值越大，则表明在第 i 种解读方式下，该 DNA 序列中各类氨基酸出现的概率更接近于氨基酸在 DNA 序列中出现概率的理论值。因此，应选取 $\max\{E_i\}$，并将对应的第 i 种解读方式作为最合理的解读方式。

$$D_{ij}=\frac{\text{属于第 } j \text{ 类氨基酸的数量}}{\text{该序列第 } i \text{ 种解读方式下氨基酸的总数}}\quad(i=1,2,3;\ j=1,\cdots,4)\tag{16}$$

具体选取解读方式并生成序列的特征向量的步骤如下：

Step 1　设 $M_j=\frac{\text{第 } j \text{ 类氨基酸所对应密码子的数目}}{64}$

$(j=1,\cdots,4)$，$\vec{P}=(M_1,M_2,M_3,M_4)$。

代入数据，$\vec{P}=\left[\frac{4}{64},\frac{10}{64},\frac{47}{64},\frac{3}{64}\right]$。

Step 2　对于一个给定的 DNA 序列，计算出每类氨基酸在第 i 种解读方式下($i=1,2,3$)在本序列中出现的频率 D_{ij}(见式(16))。然后 $D_{i1},\cdots,D_{i4}$ 同时构成一个 4 维向量 $\vec{P}_i$($i=1,2,3$ 对应 3 种不同的解读方式)。

Step 3　计算数学期望 $E_i=\vec{P}\cdot(\vec{P}_i)^T$($i=1,2,3$ 对应 3 种解读方式)。选取 $\max\{E_i\}$ 对应的第 i 种解读方式。相应的向量 $\vec{P}_i$ 即为该 DNA 序列对应的特征向量。

2.3　DNA 序列分类的 Fisher 判别

将上述构造 4 维特征向量的方法应用于文献[9]给出的训练集中 A、B 两类 DNA 序列样本，经过训练、学习，构造出以下形式的 Fisher 判别函数：$y=c_1x_1+c_2x_2+c_3x_3+c_4x_4$，其中向量 $(c_1,c_2,c_3,c_4)^T$ 即为本文“1.1”所述的向量 W。再找出 2 类样本投影的临界值，最后进行 A、B 两类的判别分类。具体步骤描述如下。

Step 1　使用本文“2.2”介绍的生成一个给定 DNA 序列特征向量的步骤，将 A 类 10 个序列得到

的10个特征向量，分别作为矩阵A中的10个列向量，从而组成以下的A矩阵。类似地，B类10个序列的特征向量组成B矩阵：

$$A=\begin{bmatrix} x_{11}^0 \cdots\cdots x_{10,1}^0 \\ x_{12}^0 \cdots\cdots x_{10,2}^0 \\ x_{13}^0 \cdots\cdots x_{10,3}^0 \\ x_{14}^0 \cdots\cdots x_{10,4}^0 \end{bmatrix} \tag{17}$$

$$B=\begin{bmatrix} x_{11}^1 \cdots\cdots x_{10,1}^1 \\ x_{12}^1 \cdots\cdots x_{10,2}^1 \\ x_{13}^1 \cdots\cdots x_{10,3}^1 \\ x_{14}^1 \cdots\cdots x_{10,4}^1 \end{bmatrix} \tag{18}$$

Step 2 分别计算出A类和B类10个样本序列的均值 m_1 和 m_2：

$$m_1=\begin{bmatrix} \bar{x}_1^0 \\ \bar{x}_2^0 \\ \bar{x}_3^0 \\ \bar{x}_4^0 \end{bmatrix} \qquad m_2=\begin{bmatrix} \bar{x}_1^1 \\ \bar{x}_2^1 \\ \bar{x}_3^1 \\ \bar{x}_4^1 \end{bmatrix} \tag{19}$$

其中，$\bar{x}_j^0=\frac{1}{10}\sum_{i=1}^{10}x_{ij}^0$，$\bar{x}_j^1=\frac{1}{10}\sum_{i=1}^{10}x_{ij}^1$ $(j=1,\cdots,4)$。

Step 3 作新的矩阵 A^*、B^*，再计算A类和B类各自的离散度矩阵 S_1 和 S_2，以及总类内离散度矩阵 S_W：

$$A^*=\begin{bmatrix} x_{11}^0-\bar{x}_1^0 \cdots\cdots x_{10,1}^0-\bar{x}_1^0 \\ x_{12}^0-\bar{x}_2^0 \cdots\cdots x_{10,2}^0-\bar{x}_2^0 \\ x_{13}^0-\bar{x}_3^0 \cdots\cdots x_{10,3}^0-\bar{x}_3^0 \\ x_{14}^0-\bar{x}_4^0 \cdots\cdots x_{10,4}^0-\bar{x}_4^0 \end{bmatrix} \tag{20}$$

$$B^*=\begin{bmatrix} x_{11}^1-\bar{x}_1^1 \cdots\cdots x_{10,1}^1-\bar{x}_1^1 \\ x_{12}^1-\bar{x}_2^1 \cdots\cdots x_{10,2}^1-\bar{x}_2^1 \\ x_{13}^1-\bar{x}_3^1 \cdots\cdots x_{10,3}^1-\bar{x}_3^1 \\ x_{14}^1-\bar{x}_4^1 \cdots\cdots x_{10,4}^1-\bar{x}_4^1 \end{bmatrix} \tag{21}$$

$$S_1=A^*(A^*)^T \tag{22}$$

$$S_2=B^*(B^*)^T \tag{23}$$

$$S_W=S_1+S_2 \tag{24}$$

Step 4 计算最优Fisher判别函数的系数 c_1，c_2，c_3，c_4：

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}=S_W^{-1}(m_1-m_2) \tag{25}$$

将所有数据代入，并将(19)式中 m_1、m_2 的值代入(25)式，求得：

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}=\begin{bmatrix} -2.312\,1 \\ -2.862\,3 \\ -0.289\,7 \\ -6.666\,7 \end{bmatrix}$$

Step 5 写出Fisher判别函数为：

$$y=-2.312\,1x_1-2.862\,3x_2-0.289\,7x_3-6.666\,7x_4 \tag{26}$$

Step 6 根据如下式子，分别计算A类和B类样本均值的判别值 y_A^* 和 y_B^*，以及临界值 y_e：

$$y_A^*=c_1\bar{x}_1^1+c_2\bar{x}_2^0+c_3\bar{x}_3^0+c_4\bar{x}_4^0 \tag{27}$$

$$y_B^*=c_1\bar{x}_1^1+c_2\bar{x}_2^1+c_3\bar{x}_3^1+c_4\bar{x}_4^1 \tag{28}$$

$$y_e=\frac{p\cdot y_A^*+q\cdot y_B^*}{p+q} \quad (p=10,q=10) \tag{29}$$

代入数据求得：$y_A^*=-0.636\,8$，$y_B^*=-1.416\,5$，$y_e=-1.026\,7$。

Step 7 作判别。例如，若有一待判别的DNA序列，其特征向量为 $(x_{01},x_{02},x_{03},x_{04})$，则可求出其判别值为 $y=-2.312\,1x_{01}-2.862\,3x_{02}-0.289\,7x_{03}-6.666\,7x_{04}$。再根据以下准则作判别：

(1)当 $y_A^*>y_e>y_B^*$ 时，若 $y>y_e$，则判别该序列属于A类；若 $y<y_e$，则判别该序列属于B类。

(2)当 $y_B^*>y_e>y_A^*$ 时，若 $y>y_e$，则判别该序列属于B类；若 $y<y_e$，则判别该序列属于A类。

3 结果与分析

3.1 实验数据来源及分类

本Fisher判别分析实验的相关数据来自文献[9]，数据分为训练集和测试集。训练集中有已知类别的DNA序列20个，编号1～10为A类，11～20为B类。测试集数据共2组，第1组是一个小样本集合，有编号为21～40的20个未知类别的DNA序列样本；第2组是一个大样本集合，共182个DNA序列。

本实验包括对测试集序列样本分类实验和检测Fisher判别分类准确度的实验。实验环境采用CPU为奔腾双核2.60 GHz，2.00 GB内存，Microsoft Windows XP (SP2)操作系统，Matlab 7.0进行编程设计[10]。

3.2 分类实验的结果

利用训练集中的样本序列，即可得到本文“2.3”所述Fisher判别函数(见式(26))。应用该判别函数，对测试集中编号为21～40的DNA序列进行分类，并将分类结果与人工神经网络(ANN)方法[2]、支持向量机(SVM)方法[3]的分类结果进行对比(表3)。

由表3可知，本文方法的分类结果与ANN方法的结果完全一致，与SVM方法的结果也基本相同。但是，本文方法从概率统计的角度出发，与其他方法相比，具有理论性强、判别速度快等优点。

表 3　不同分类方法的结果对比

Table 3　Comparison of results based on different classification methods

分类结果 Classification results	SVM 方法 SVM method	ANN 方法 ANN method	本文 Fisher 方法 Fisher method
A 类 DNA 序列的编号 Numbers of DNA sequence of A class	23, 25, 27, 29, 34, 35, 37	22, 23, 25, 27, 29, 34, 35,37,39	22, 23, 25, 27, 29, 34, 35,37,39
B 类 DNA 序列的编号 Numbers of DNA sequence of B class	21, 22, 24, 26, 28, 30, 31, 32, 33, 36, 38, 39, 40	21, 24, 26, 28, 30, 31, 32, 33, 36, 38,40	21, 24, 26, 28, 30, 31, 32, 33, 36, 38,40

3.3　分类准确度检验

根据本文“1.2”介绍的回代估计法，应用本文建立的 Fisher 判别函数，对训练集中 20 个已知类别的 DNA 序列样本进行回代检测。回判结果为 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1，其中 0 代表 A 类样本，1 代表 B 类样本。回判结果与实际情况比较，并代入回判公式得到误判率为 0，即分类正确率为 100%。

进一步按照以下步骤对大样本集中 182 个序列样本进行回代检测：

(1)利用前面得到的 Fisher 判别函数对 182 个样本进行分类。

(2)将得到的 182 个已分类的样本作为训练集，使用上述的 Fisher 判别法再次产生判别函数。

(3)用再次得到的判别函数对这 182 个样本进行回代检测。将回判结果与 182 个样本初始的分类结果进行比较，代入回判公式得到误判率为2.49%，即分类正确率为 97.51%。

综合以上回判检测的结果，说明本文 Fisher 判别法构建的判别函数拥有很好的分类正确率，本方法可以作为判别 DNA 序列类别的方法。

但是由于 Fisher 判别法的判别函数是线性函数，对该方法的实用造成一定的局限性，因此今后还需对它作进一步的研究和改进。

参　考　文　献

[1] 徐晓秋，初立元，左铭杰，等. DNA 分类方法的探讨[J]. 大连大学学报，2001，22(4)：95-100.

[2] 冯涛，康雯，韩小军，等. 关于 DNA 序列分类的模型[J]. 数学的实践与认识，2001，31(1)：26-30.

[3] 徐健，李柏年，张孔生，等. 基于 SVM 分类机的一种 DNA 序列判别方法[J]. 安徽理工大学学报：自然科学版，2009(3)：58-61.

[4] 王学武. 眼睛梯度特征的人脸检测及 Fisher 人脸识别技术的应用[D]. 湘潭：湘潭大学图书馆，2006.

[5] 范金城，梅长林. 数据分析[M]. 北京：科学出版社，2002：175-200.

[6] ATIYAH M. Mathematics：frontiers and perspectives [M]. Providence：AMS，2000：43-50.

[7] 周玉元，周铁军. DNA 序列分类的 Fisher 判别法[J]. 湖南农业大学学报：自然科学版，2003(10)：438-440.

[8] 乔明艳，李全斌. 对生物化学中氨基酸分类有关问题的讨论[J]. 卫生职业教育，2006，24(23)：153-154.

[9] 中国工业与应用数学学会. 2000 年全国大学生数学建模竞赛 A 题[EB/OL]. [2012-09-20] http://www.mcm.edu.cn/2000.

[10] 胡守信，李柏年. 基于 matlab 的数学实验[M]. 北京：科学出版社，2004：181-189.

A classification method of DNA sequence based on Fisher discriminant analysis

YANG Li-ping[1]　LU Song-feng[2]　HU He-ping[2]　HUANG Yu[1]

1. *College of Sciences*, *Huazhong Agricultural University*, *Wuhan* 430070, *China*;
2. *College of Computer Science and Technology*, *Huazhong University of Science and Technology*, *Wuhan* 430074, *China*

Abstract　A new classification method of DNA sequence based on Fisher discriminant analysis was proposed. According to the properties of amino acid, a feature vector space corresponding to DNA sequences was established. Then features of two types DNA sequences in the training set were extracted, and a Fisher discriminant function was established. This function was applied to classify 182 DNA sequences in the test set. The results showed that the proposed method has excellent classification accuracy.

Key words　bioinformatics; Fisher discriminant analysis(FDA); classification of DNA sequence; feature extraction; feature vector space

(责任编辑：边书京)