任意选出比较多的(为了保证较高的准确性), 利用 keyword 作为分类标准, 然后利用本文提供的加权系数的确定方法就可以定出一个具体的定量标准 具有一定实用价值

参考文献:

[1] 李 涛, 贺勇军等. MATLAB 工具箱应用指南——应用数学篇. 电子工业出版社.
[2] 袁亚湘. 最优化方法. 科学出版社.
[3] 张乃孝, 裘宗燕. 数据结构——c+ + 与面向对象的途径. 高教出版社.
[4] 汪仁官. 概率论引论. 北京大学出版社.
[5] 陈家鼎, 孙山泽等. 数理统计学讲义. 高教出版社.

The Grouping of DNA Sequences Model

YANG Jian, WANG Chi, YANG Yong

(Peking University, Beijing 100871)

Abstract: In this paper, a method to classify the DNA sequences is proposed. Mathematical methods such as statistics and optimization are used to build the model. The data is analysed sufficiently and the "critical words" is got, which can represent the characteristics of each group. According to this, a quantitative standard for grouping is brought forward. This model can properly classify the given data through testing. First, the strings which appear repeatedly (called words) in the given data are scanned out. The standard frequency and dispersion for each word are calculated. Second, using the Least Squares method, the priority function is fixed. Through stepwise optimization, the coefficients are made stable. Third, the key words are selected out and calculate the weight according to the priority function. At last, using the "analyse hierarchy process", the undetermined data is classified. This method can classify the undetermined data (No. 21—No. 40) fairly well, it can also give good result for the last 182 sequences.

DNA 序列的分类

韩铁平, 余 杭, 刘 威
指导老师: 杨启帆

(浙江大学, 杭州 310027)

编者按: 本文借助于计算机符号处理的能力来把握序列中不同碱基的丰度特征, 从而进行了利用数理统计方法的分类研究. 而后引入相关度分类判别算法及反馈机制来比较碱基的相对位置, 在既定方向上颇具新意地把工作推向深入. 不足之处在于, 未能使用相关度工具对各类样本分别进行分析; 此外, "纯数学"必须与其他学科紧密结合才会有优秀的建模工作, 本文虽然对编码氨基酸的三联体进行初步探讨, 着墨处自是轻淡许多.

摘要:　本文对A 题中给出的DNA 序列分类问题进行了讨论　从"不同序列中碱基含量不同"入手建立了欧氏距离判别模型, 马氏距离判别模型以及Fisher 准则判定模型; 又从"不同序列中碱基位置不同"入手建立了利用序列相关知识的相关度分类判别算法, 并进一步研究了带反馈的相关度分类判别算法　对于题中所给的待分类的人工序列和自然序列, 本文都一一作了分类　接着, 本文又对其它各种常见的分类算法进行了讨论, 并着重从分类算法的稳定性上对几种方法作了比较

1　问题的重述(略)

2　模型的条件和假设(略)

3　符号约定

na: 任一给定序列中碱基A 的百分含量;
ng: 任一给定序列中碱基G 的百分含量;
nt: 任一给定序列中碱基T 的百分含量;
nc: 任一给定序列中碱基C 的百分含量
Gi: 由某些具有相同属性的个体组成的类

4　问题的分析和解答

4.1　概述

根据题意, 我们首先要提取出一个序列的特征, 然后给出它的数学表示, 最后选择并构造基于这种数学表示的分类方法　对于一个任意一个DNA 序列, 我们认为, 反映该序列特征的方面有两个:

1. 碱基的含量, 反映了该序列的内容;
2. 碱基的排列情况, 反映了该序列的形式

4.2　基于碱基含量特征分类的模型

首先, 我们考虑采用序列中的A, G, T, C 的含量百分比作为该序列的特征　这样的抽取特征的方法具有其生物学的意义　前面提到过, 在不用于编码蛋白质的序列片断中,A 和T 的含量特别多些, 因此以某些碱基特别丰富作为特征去研究DNA 序列的结构是具有可行性的　将序列中的A, G, T, C 的含量百分比分别记为na, ng, nt, nc, 则得到一组表征该序列特征的四维向量(na, ng, nt, nc).　考虑到na, nt, ng, nc 线性相关$(na+ ng+ nt+ nc= 1)$, 所以我们采用简化的三维向量(na, nt, ng)来进行计算　对于标号为i 的序列, 记它的特征向量为X_i.　显然, 任意序列的特征向量与一个3 维空间的点对映

一般的判别问题为: 设有k 个类别$G_1, G_2, \cdots, G_k$, 对任意一个属于G_i 类样品x, 其特征向量X 的值都可以获得　现给定一个由已知类别的一些样品$x_1, x_2, \cdots, x_n$ 组成的学习样本, 要求对一个来自这k 个类别的某样品x, 根据其特征向量X 的值作出其所属类别的判断

在本题DNA 序列分类中, $k= 2, G_1= A, G_2= B$, 特征向量X 是三维的　学习样本共包含$n= 20$ 个样本, 其中10 个属于A, 10 个属于B.　我们分别采用了欧氏距离(Euclid)分类模型, 马氏距离(M ahalanobis)分类模型和Fisher 判别模型来对序列样本分类

4.2.1 欧氏距离(Euclid)分类模型

在欧氏距离(Euclid)分类模型中,把每个样本视为三维空间的一个点,以其到不同集合几何中心的欧氏距离作为判据 具体的算法如下:

1. 计算属于A 类与属于B 类的10 个样本点的集合各自的几何中心:

$$C_A = \frac{1}{10}\sum_{i=1}^{10} X_i \qquad C_B = \frac{1}{10}\sum_{i=11}^{20} X_i$$

2. 对于给定的样本点X_i,分别计算该点到C_A 的欧氏距离$D_A = |X_i - C_A|$,以及该点到C_B 的欧氏距离$D_B = |X_i - C_B|$;

3. 判别准则如下:

(1) 若$D_A < D_B$,则将X_i 点判为A 类;

(2) 若$D_A > D_B$,则将X_i 点判为B 类;

(3) 若$D_A = D_B$,则将X_i 点判为不可判类;

用上述算法对已知样学习样本A 1—A 20 进行分类,结果是除了A 4 被错误的分到了B 类外,其余的19 个样本全部正确,分类准确率达到95%.

用上述算法对未知的人工序列A 21—A 40 进行分类,得到的结果是:

A 类: 22, 23, 25, 27, 29, 30, 32, 34, 35, 36, 37, 39; B 类: 21, 24, 26, 28, 31, 33, 38, 40

用上述算法对未知的自然序列N 1—N 182 进行分类,得到的结果见附录 (略)

用欧氏距离作为判据虽然简便直观,但存在着明显的缺陷:从概率统计的角度来看,用欧氏距离描述随机点之间的距离并不好. 因此当待分类样本是随机样本,具有一定的统计性质时,这个模型并不能很好的描述两个随机点之间的接近程度

4.2.2 马氏距离(Mahalanobis)分类模型

为了克服采用欧氏距离时的缺陷,我们采用马氏距离来代替欧氏距离 改进后的算法如下:

设:三维总体G 的均值为$\mu = (\mu_1, \mu_2, \mu_3)^T$,协方差矩阵为非奇异阵$V_{3\times 3}$, 则三维样本$X$ 到总体G 的马氏距离为:

$$dm(X, G) = \sqrt{(X-\mu)^T V^{-1}(X-\mu)}$$

其中未知的μ 可用学习样本的均值来代替,协方差矩阵V 可用学习样本的样本协方差矩阵来代替.

将马氏距离用于判别模型,遵循判据如下:

1. 若$dm(X, A) < dm(X, B)$,则判定x 为A 类;

2. 若$dm(X, A) > dm(X, B)$,则判定x 为B 类;

3. 若$dm(X, A) = dm(X, B)$,则判定x 为不可判类;

用上述算法对已知样学习样本A 1—A 20 进行分类,结果是除了A 4 被错误的分到了B 类外,其余的19 个样本全部正确,分类准确率达到95%.

用上述算法对未知序列A 21—A 40 进行分类,得到的结果是:

A 类: 22, 23, 25, 27, 29, 30, 32, 33, 34, 35, 36, 37

B 类: 21, 24, 26, 28, 31, 38, 39, 40

用上述算法对未知的自然序列N 1—N 182 进行分类,得到的结果见附录. (略)

4.2.3　Fisher 准则分类模型

在多维空间里分类的方法不仅仅是距离分类法一种, 常用的 Fisher 分类法就是另一种基于几何特性的分类法　在距离判别模型中, 三维空间的样品 X 被映射为一维的距离 d 来作判断　Fisher 分类法的思想也是把三维空间的样本映射为一维的特征值 y, 并依据 y 来进行判别　具体的作法是先引入一个与样本同维的待定向量 u, 再将 y 取为 X 坐标的线性组合 $y = u^T x$.　而 u 的选取　要使同一类别产生的 y 尽量聚拢, 不同类别产生的 y 尽量拉开. 这样, 我们便可将样品 X 到某一类 G 的距离定义为 $y = u^T x$ 与 $yc = u^T c$ 之间的欧氏距离:

$$L(X, G) = |y - yc| = |u^T (x - c)|$$

其中 c 为 G 的几何中心.

Fisher 分类的判据为:

1 若 $L(X, A) < L(X, B)$, 则判定 x 为 A 类;

2 若 $L(X, A) > L(X, B)$, 则判定 x 为 B 类;

3 若 $L(X, A) = L(X, B)$, 则判定 x 为不可判类

根据对 u 的要求, Fisher 提出了比较有效的选择算法, 利用该算法, 从学习样本中获得:

$$u = (0.3365, -0.087, 0.9377)^T$$

$$L(X, A) = |0.3365^* (na - 0.2860) - 0.087^* (nt - 0.1550) + 0.9377^* (ng - 0.3830)|$$

$$L(X, B) = |0.3365^* (na - 0.2940) - 0.087^* (nt - 0.5010) + 0.9377^* (ng - 0.1010)|$$

用上述算法对已知样学习样本 A1—A20 进行分类, 结果仍然是除了 A4 被错误的分到了 B 类外, 其余的 19 个样本全部正确, 分类准确率达到 95%.

对于未知序列 A21—A40 进行分类, 得到的结果是:

A 类: 22, 23, 25, 27, 29, 34, 35, 36, 37; B 类: 21, 24, 26, 28, 30, 31, 32, 33, 38, 39, 40

用上述算法对未知的自然序列 N1- N182 进行分类, 得到的结果见附录 . (略)

4.2.4　三种距离分类模型的比较

这三种模型在分类结果上有一定的区别, 对于序列 A30, A32, A33 及 A39, 三种方法给出了不同结果, 见表 1.

表 1

	欧氏距离法	马氏距离法	Fisher 准则法
30	A	A	B
32	A	A	B
33	B	A	B
39	A	B	B

对于这种情况, 我们提出一个联合判定准则: 对于任一个序列, 当三种分类法结果完全一致时, 认为它判别有效; 若不然, 当三种分类法结果不一致时, 认为该序列为不可判类

对于三种方法都无法正确分类的 A4 序列, 可认为是异常情况, 不影响算法的性能

4.3　基于碱基位置特征分类的模型

虽然上述采用碱基 A, T, G, C 在 DNA 序列里的含量作为该序列的特征的方法有一定的生物学意义并且在 DNA 序列的分类中获得了比较理想的结果　但是, 用这种方法抽取特征, 没有充分体现碱基排列的信息量, 仅仅考虑碱基含量并没有体现碱基在序列中的排列情况　例如, 序列 (ATGC) 与序列 (CGTA) 有着相同的碱基含量, 他们的特征向量是完全一样的, 并不能体现在排列结构上的不同　因此, 直接从序列本身的碱基排列顺序来考察序列就成为一种更加合适的提取特征的方式　因此采纳数值序列中的相关性分析设计了算法

通常任意两个数值序列的相关性都是通过这两个序列的相关函数来刻画的. 由于本题中的DNA 序列是非数值的序列, 同时无法将碱基按通常的方式进行数值化, 因而刻画任意两个序列的相关程度的变量需要重新定义

4.3.1 定义一: 相关运算"⊙"

对于任意碱基 m 和 n, 相关运算"$m \odot n$"的值由表 2 定义:

表 2

⊙	A	G	T	C
A	1	0	0	0
G	0	1	0	0
T	0	0	1	0
C	0	0	0	1

4.3.2 定义二: 哑元O

除四个碱基外, 我们另行定义一个哑元 O, 规定任意碱基与哑元作相关运算的结果都为 0.

4.3.3 定义三: 序列的延拓

对于任意一个长度为 N 的序列 A_i (其中 $0 \leq i < N$), 定义它的延拓为如下一个无限序列:

A^+j: 当 $0 \leq j < N$ 时, $A^+_j = A_j$; 当 $-\infty < j < 0$ 及 $N \leq j < \infty$ 时, $A^+j = O$.

即在该序列的左右两端均用哑元 O 填充

4.3.4 定义四: 序列的相关度

对于任意的两个序列 A_N, B_M, 定义序列 A 和序列 B 的相关序列 S_i 为:

$$S_i = \sum_{k=0}^{\infty} A^+_{k+2-i} \odot B^+_k \quad (0 \leq i \leq n+m-1)$$

定义序列 B 对序列 A 的相关度为:

$$S = MAX\{S_i\} \quad (0 \leq i \leq n+m-1)$$

例如对于序列 $A\{T, C, T\}$ 与序列 $B\{A, G, T, C, T, C\}$, 相关序列及相关度的计算步骤如下:

第一项: $S_0 = A_2 \odot B_0 = T \odot A = 0$

…	A^+_{-1}	A^+_0	A^+_1	A^+_2	A^+_3	A^+_4	A^+_5	A^+_6	A^+_7	A^+_8	…
…	O	T	C	T	O	O	O	O	O	O	…
…	O	O	O	A	G	T	C	T	C	O	…
…	B^+_{-3}	B^+_{-2}	B^+_{-1}	B^+_0	B^+_1	B^+_2	B^+_3	B^+_4	B^+_5	B^+_6	…

第二项: $S_1 = A_1 \odot B_0 + A_2 \odot B_1 = T \odot G + C \odot A = 0$

…	A^+_{-2}	A^+_{-1}	A^+_0	A^+_1	A^+_2	A^+_3	A^+_4	A^+_5	A^+_6	A^+_7	…
…	O	O	T	C	T	O	O	O	O	O	…
…	O	O	O	A	G	T	C	T	C	O	…
…	B^+_{-3}	B^+_{-2}	B^+_{-1}	B^+_0	B^+_1	B^+_2	B^+_3	B^+_4	B^+_5	B^+_6	…

第三项: $S_2 = A_0 \odot B_0 + A_1 \odot B_1 = T \odot T + G \odot C + A \odot T = 1$

…	A^+_{-3}	A^+_{-2}	A^+_{-1}	A^+_0	A^+_1	A^+_2	A^+_3	A^+_4	A^+_5	A^+_6	…
…	O	O	O	T	C	T	O	O	O	O	…
…	O	O	O	A	G	T	C	T	C	O	…
…	B^+_{-3}	B^+_{-2}	B^+_{-1}	B^+_0	B^+_1	B^+_2	B^+_3	B^+_4	B^+_5	B^+_6	…

以下类推得(表略):

第四项: $S_3 = A_0 \odot B_1 + A_1 \odot B_2 + A_2 \odot B_3 = T \odot C + C \odot T + T \odot G = 0$

第五项: $S_4 = A_0 \otimes B_2 + A_1 \otimes B_3 + A_2 \otimes B_4 = T \otimes T + C \otimes C + T \otimes T = 3$

第六项: $S_5 = A_0 \otimes B_3 + A_1 \otimes B_4 + A_2 \otimes B_5 = T \otimes C + C \otimes T + T \otimes C = 0$

第七项: $S_6 = A_0 \otimes B_4 + A_1 \otimes B_5 = C \otimes C + T \otimes T = 2$

第八项: $S_7 = A_0 \otimes B_5 = T \otimes C = 0$

第八项: $S_7 = A_0 \otimes B_5 = T \otimes C = 0$										
…	A^+_8	A^+_7	A^+_6	A^+_5	A^+_4	A^+_3	A^+_2	A^+_1	A^+_0	A^+_1	…
…	*O*	*O*	*O*	*O*	*O*	*O*	*O*	*O*	T	*C*	…
…	*O*	*O*	*O*	*A*	*G*	*T*	*C*	*T*	C	*O*	…
…	B^+_3	B^+_2	B^+_1	B^+_0	B^+_1	B^+_2	B^+_3	B^+_4	**B**$^+_5$	B^+_6	…

两序列的相关度为 $S = MAX\{S_i\} = S_5 = 3$;

4.3.5　定理一: 任意给定三个序列 S, A, B, **若** A **与** S **的相关度大于** B **与** S **的相关度且** B **与** A **等长, 则** A **与** S **属同一类的可能性大于** B **与** S **属同一类的可能性**

4.3.6　基于相关度的分类算法:

利用上述概念, 我们构造了一个基于相关度的分类算法, 如下:

1. **对于序列** A 21—A 40, N 1—N 182 **中的任意一个序列, 将其与序列** A 1—A 20 **中的每一个依次作求相关度的运算, 结果记为** SS1, SS2, SS3……SS20;

2. **对于前十个相关度, 求出它们的平均相关度** SA = (SS1+ SS2+ ……SS10)/10, **并定义其为与** A **类的相关度;**

3. **对于后十个相关度, 求出它们的平均相关度** SB = (SS11+ SS12+ ……SS20)/10, **并定义其为与** B **类列的相关度;**

4. **记** W = SA/SB, **根据定理一, 判别依据为:**

若 W > 1, **则将** X **点判为** A **类;**

若 W < 1, **则将** X **点判为** B **类;**

若 W = 1, **则将** X **点判为不可判类;**

5. W **可作为衡量该序列分类的可信性的一个标准　显然当** W **越接近于** 1, **该序列与** A **类的相关性和与** B **类的相关性区别就越小, 分类结果就越不可信; 反之,** W **与** 1 **差的越远, 该序列与** A **类的相关性和与** B **类的相关性区别就越小, 分类结果就越可信　这个变量对我们下面带有反馈的相关度分类算法具有重要的意义**

用上述算法对已知样学习样本 A 1- A 20 **进行分类, 得到的结果是分类完全正确,** A, B **类可以完全分开, 准确率达到** 100%.

对于未知序列 A 21—A 40 **进行分类, 得到的结果是:**

A **类:** 22 23 25 27 29 34 35 36 37

B **类:** 21 24 26 28 30 31 32 33 38 39 40

用上述算法对未知的自然序列 N 1—N 182 **进行分类, 得到的结果见附录(略).**

4.3.7　相关度分类算法的改进——带有反馈的分类算法

上述的相关度分类算法是一次性学习过程, 学习的过程只体现在学习样本的过程中, 而在对未知样本分类的过程中没有对已分类情况作出修正, 即是属于无反馈型的学习　然而, 采用反馈型的学习过程会有更好的分类结果　一般说来, 带反馈的算法以神经网络算法最具有代表性　但对于一般的分类算法而言, 可以采用多次反复分类的办法来实现反馈的目

的. 针对上述的相关度分类算法, 我们设计了如下带反馈的相关度分类算法:

1. 对全部 182 个样本进行相关度分类;

2. 计算全部 182 个W 的值

3. 在所有被判为A 类的待分类序列中, 取出W 值最大的一个, 作为标准学习样本, 加入到A 类的标准样本中(若有多个, 则全部加入到A 类中, 若无被判为A 类的序列, 则保持A 类标准学习样本不变)

4. 在所有被判为B 类的待分类序列中, 取出W 值最小的一个, 作为标准学习样本, 加入到B 类的标准样本中(若有多个, 则全部加入到B 类中, 若无被判为B 类的序列, 则保持B 类标准学习样本不变)

5. 重复对剩余的待分类序列进行相关度分类, 并按上述步骤不断扩充标准学习样本, 直至全部的待分类序列都被加入到标准学习样本中.

我们用新算法编程对 182 个序列进行了重新分类, 得到了不同于原无反馈分类算法的结果, 而且新的分类结果的W 值明显与 1 离开的更大, 这使我们有理由相信, 反馈对算法的性能有一定的改进

5 进一步研究的问题

5.1 基于生物学的特征抽取

我们上述的两种特征抽取方法更多的是从纯数学眼光来研究序列的特征. 除此之外, 我们还可以考虑DNA 序列在生物学意义下的数学特征

一个比较容易考虑到的方面便是三联体在DNA 序列中的出现. 由于具有三联体形式的遗传密码子对蛋白质的合成具有决定性作用, 有理由认为它在序列中的出现体现了该序列的本质特征. 题中没有明确的指明所给的序列是全序列还是序列片断, 我们无法对三联体在序列中的出现位置进行定位, 一种代替的方法是将序列假定为全序列, 从第一个碱基开始三个三个一组的划分为密码子, 然后统计 64 个密码子的出现概率, 形成 64 维的向量. 再使用距离分类等模型, 或利用生物学的知识先将 64 维向量的某几维合并, 降维后再分类. 我们编程演算后, 觉得该种分类方法比较依赖于密码子的划分, 一位碱基的缺失或错位均会造成分类错误, 所以必须加以修改, 一条思路是尝试将序列移一位或二位再划分密码子, 由于时间所限, 没有进一步研究

5.2 基于人工神经网络的模型

人工神经网络是一种带反馈的自适应算法, 随着计算机速度提高被广泛应用. 对于本题的情况采用神经网络模型是合适的, 它可以在给定特征向量的情况下代替一般的距离分类模型. 对于基于碱基含量的特征向量(na, nt, ng), 构造了如下的反向传播算法:

1. 网络简单的分为两层, 一层为输入层, 有 3 个单元, 分别为权重 a, b, c; 一层为输出层, 有 1 个单元, 为判别结果; 各单元均为 Sigmoid 型函数激励

2. 设定(a, b, c)的初值为$(0, 0, 0)$; A 类学习样本的标准输出定为 1; B 类学习样本的标准输出定为 0

3. 对每一个学习样本, 计算$S = a * na + b * nt + c * ng$ 作为输出;

4. 将学习样本的标准输出与S 相减, 所得的差用来指导权重的改变, 权重的改变遵从 Widrow-Hoff 准则

5. 反复学习样本, 到权重值稳定收敛

6. 代入待分类样本, 分类

用上述算法所得到的结果与普通的分类模型没有区别　事实上当权值稳定收敛后, $S = a * na + b * nt + c * ng$ 就是特征空间的一张(超)平面, 从这一点来说, 人工神经网络模型与一般的距离分类模型得到的结果没有两样　考虑到人工神经网络模型还存在结果对初值有较强敏感性, 缺乏选择理想步长的准则和收敛性等问题, 在一定的时间内, 我们无法较好的解决这些问题, 所以我们也没有作进一步讨论

6　算法的稳定性

前面比较算法的时候, 曾多次提到分类算法的稳定性问题　分类算法的稳定性是除了算法的成功率之外的另一较重要的指标　所谓分类算法的稳定性, 是指算法在样本发生了轻微变化时作出正确判别的能力　对于本题, 是指算法在样本序列发生了轻微的碱基缺失, 错位, 错排情况时作出正确判别的能力　因为本题要求我们研究的是DNA 序列粗粒化和模型化的问题, 所以分类时是对序列的整体特征进行区分　局部碱基的组成变化应该对算法的分类结果没有影响　我们所提出的几个模型均较好的满足了这一点

参考文献:

[1] 孙乃恩, 孙东旭, 朱德熙 《分子遗传学》. 南京大学出版社, 1996.

[2] 白其峥 .《数学建模案例分析》. 海洋出版社, 2000.

[3] 潘德惠 .《数学模型的统计方法》. 辽宁科学技术出版社, 1986.

[4] 阎平凡, 黄端旭 .《人工神经网络》. 安徽教育出版社, 1991.

[5] 李振刚 .《分子遗传学概论》. 中国科学技术大学出版社, 1990.

[6] Duane Hanselman·Bruce Littlefield《Mastering MATLAB: a comprehensive tutorial and reference》. Prentice Hall, 1996.

Classification of DNA Sequences

HAN Yi-ping,　YU Hang,　LIU Wei

(Zhejiang Univ., Hangzhou 310027)

Abstract: This paper proposes several methods for the classification of DNA sequences. We noticed that different sequences have different alkali radicals and therefore set up models using Euclidean distance, Mahalanobis distance and Fisher principle. We also noticed that different sequences have different permutations of alkali radicals and an algorithm using relativity analysis is proposed. Further we discussed a relativity analysis algorithm with feed-back mechanism. As to the natural and artificial data given our algorithms work well and fine results are given. At last several other common algorithms are compared, especially on their stabilities.