知识介绍

细菌的数值分类方法

郭秀花 李爱国 冯 丹 张德舜 朴相浩

(解放军北京医学高等专科学校，北京 100071)

数值分类已广泛应用于微生物学、动物学、植物学、古生物学等领域[1]。本文就细菌数值分类进行讨论。

1 数值分类的概念及发展概况

细菌分类方法有：传统分类法、数值分类法、分子遗传学分类法、化学分类法等[2]。Sokal 和 Sneath 指出数值分类是：借助数值方法，根据其性状、状态将分类单位归类成类元。我国的周方和陈宁庆也有过通俗的解释[3]。数值分类的实现过程需借助计算机，因此，细菌的数值分类就是将数学分类的理论和方法，借助于计算机，应用到细菌方面的一门学科。

数值分类学最早可追溯到上个世纪生物度量学的兴起。早在 1898 年，Heickle 使用了距离测度区分鲱鱼种系，1957～1961 年建立数值分类学的最初理论和方法。1963 年 Sokal 和 Sneath 出版了《数据分类学原理》一书，从此，数值分类学迅猛发展。1973 年 Sokal 和 Sneath 对国际上常用算法的数学模型进行了系统归纳[1]。1984 年 Sneath 将数值分类列为细菌分类方法之首[4]，概述了数值分类的程序及要领，并且肯定，用相似性分析将细菌株排列成的表元是广泛地等同于类元的。1988 年 Dymbowski Franklin 和 Lepage 首先使用计算机进行肠道菌分类项目的概率运算，并证明其可靠性[5]。在我国，1974 年徐浩发表了“微生物数值分类法简介”[6]，并于 1975 年首次成功地将数值分类法用于 63 株枯草杆菌的研究[7]。马俊才等从 80 年代起，完成了中国微生物资源数据库（MRDC）的构建工作[8,9]，MRDC 数值库中含有数据分类软件包 MINTS。何道生、李虹等进行了模糊数学在微生物分类方面的尝试[10,11]。

2 数值分类的步骤

对 n 个菌株进行数值分类可按如下步骤进行：

(1) 确定性状指标。选取此部分菌株一些主要的性状指标。

(2) 状态测定及数据收集。若我们已确定了 m 个性状，通过正式实验，可得到每个菌株在各项性状方面试验的结果。

(3) 建立原始性状编码数据矩阵。性状编码就是把实验中测得的性状状态的记录结果转化为计算机能识别的符号，得到原始性状编码数据矩阵 $(X_{ij})_{n\times m}$。

(4) 必要时对数据进行标准化。

(5) 计算相似性系数，得相似系数矩阵或距离矩阵。

(6) 选用适当的聚类方法进行聚类。

(7) 绘树状图，得数值分类结果。

(8) 对数值分类结果进行修定。可选用几种聚类方法来进行聚类。然后结合专业进行分析，或通过对未知菌株进行鉴定，以判断聚类结果的科学性、合理性和实用性。修定了数值分类结果后，就可利用分类结果对大量的未知菌株进行鉴定。数值鉴定的方法国内外有许多报道，大部分数值鉴定的理论都是依据 Bayes 理论[12～15]，有的还编制了计算机程序[5,16,17]。

3 数值分类法

数值分类方法的不同关键在于相似性系数的计算方法和聚类方法的选取。

3.1 系统聚类法

3.1.1 编码方法：为了便于系统聚类，可将常见性状按如下方法编码：

1994-07-18收稿

(1) 二态性状：阳性阴性结果分别用 1 和 0 表示。资料缺失用 NC(不比较符号)表示，(输入计算机时用 3 或 4 代表)。

(2) 定量多态性状(或称有序多态性状)：因各状态之间有逻辑上的次序关系，如细菌细胞的长度，根瘤菌在不同浓度 NaCl 的培养基上的生长性状，对这类性状可以采用加权递增编码法谓之加权[18]，递增编码如表 1。

(3) 定性多态性状(或称无序多态性状)：一般也是转换为多种二态性状。例如表 2。

另外，有时也用表 3 形式的编码，它反映的是石蕊牛奶反应的情况。

表 1　递增编码法

菌株	性状			含义
	NaCl(1%)	NaCl(2%)	NaCl(3%)	
1	0	0	0	在 1%NaCl 上不生长
2	1	0	0	在 1%NaCl 上生长
3	1	1	0	在 2%NaCl 上生长
4	1	1	1	在 3%NaCl 上生长

表 2　加权非递增编码

菌株	结构的颜色	二态性状		
		1	2	3
1	红	1	0	0
2	黄	0	1	0
3	蓝	0	0	1

表 3　加权递增与非递增混合编码

菌株	石蕊牛奶反应				
	产酸	酸凝	产碱	胨化	还原
1	1	1	NC	1	0
2	0	NC	1	1	0
3	NC	NC	NC	1	1

李多川等采用类似方法编码对镰刀菌进行数值分类[19]。马俊才[20]在对 116 株放线菌进行数值分类中，起初确定了 75 项性状。利用性状消减，在两态性状中，去掉了那些 90%以上的菌株都呈一种状态的性状，共去掉 14 项。在形态方面的多态性状中，以每项性状的最大状态数为其权数的原则加权，使加权后整个性状数达到 95 项。

3.1.2 计算相似性系数的方法：相似性系数可分为结合系数、相关系数、距离系数及概率相似性系数等。应用最多的是结合系数。它们适于二态性状数据。

a, b, c, d 的含义见表 4。

表 4　字母含义

菌株 A	菌株 B		
	编码	1	0
编	1	a	b
码	0	c	d

a：表示两个菌株(或称OTU)的性状编码皆为1的个数，称为正匹配

b：表示菌株A的性状编码为1，而菌株B的编码为0的性状个数，称为错匹配

c：表示菌株A的性状编码为0，而菌株B的编码为1的性状个数，称为错匹配

d：表示两种性状编码皆为0，称为负匹配

Avstin 通过 36 种相似性系数采用平均连锁法进行数值分类[21]，发现有 15 种系数适合于细菌分类，其中如：

$$S_h = \frac{a+d-b-c}{a+b+c+d};\quad S_{TD} = \frac{b+c}{a+b+c+d};$$

$$S_{SM} = \frac{a+d}{a+b+c+d};\quad S_D = \frac{2a}{a+b+c+d}。$$

马俊才等给出了 8 种相似性系数[22]：

(1) $S=(a+d)/(a+b+c+d-NC)$;

(2) $S=a/(a+b+c)$;

(3) $S=2a/(2a+b+c)$;

(4) $S=a/\sqrt{(a+b)(a+c)}$;

(5) $S=b+c/\sqrt{(a+b+c+d)}$;

(6) $S=(a+d)/(a+b+c+d)$;

(7) $S=a/(a+b+c+d-NC)$;

(8) $S=1-[\sqrt{(b*c)}/(a+b+c+d-NC)]$;

7 种聚类方法为：(1)单连锁法；(2)全连锁法；(3)平均连锁法；(4)中线法；(5)形心法；(6)平方和增量法；(7)可变法。

Lance 和 Williams 1967 年确定了一个适用于七种方法的通用公式:

$$D_{kr}=A*D_{ki}+B*D_{kj}+C*D_{ij}+D*D_{ki}-D_{kj}$$

式中 D_{kr} 样品组 k 与样品组 r 间的距离,样品组 r 是样品组 i 与样品组 j 的并组(即 $G=G(\bigcup)G$)。A、B、C、D 是随不同聚合策略所取的不同系数。

另外,有不少人利用下面的符号系数采用平均连锁法进行数值分类[27],它是由 Sokal 和 Michener 二氏提出的。

$$S_{sm}=\frac{\text{阳性的阴性的符合数总和}}{\text{总测定数减无效测定数}}=\frac{(\sum+)+(\sum-)}{M-(NC)}$$

即 $$S_{sm}=\frac{a+d}{a+b+c+d-(\sum NC)}$$

3.2 模糊聚类分析

贺仲雄等[28]提出了 15 种相似性系数的计算方法。首先将各代表的性状指标的数据标准化,公式: $X=(X'-\overline{X'})/C$,其中 X': 原始数据; $\overline{X'}$: 原始数据的平均值; C: 原始数据的标准差。再将标准化数据压缩在 [0, 1] 闭区间内。

相似性系数的计算方法有:(1)闵可夫斯基距离法 3(其中含有绝对距离,欧氏距离和切比雪夫距离);(2)马氏距离法;(3)兰氏距离法;(4)数量积法;(5)相关系数法;(6)指数相似系数法;(7)夹角余弦法;(8)非参数法;(9)绝对值指数法;(10)绝对值倒数法;(11)绝对值减数法;(12)最大、最小法;(13)算数平均最小法;(14)几何平均最小法;(15)专家评分法等。

在细菌分类方面,李虹建立了一种新的模糊数学处理相似性系数计算公式:

$$r_{ij}=\frac{\sum(X_{ik}\oplus X_{jk})}{m}$$

$$(i=1,2,\cdots n, j=1,2,\cdots n)$$

其中, $$X_{ij}\oplus X_{jk}=\begin{cases}1 & X_{ik}=X_{jk}\\ 0 & X_{ik}\neq X_{jk}\end{cases}$$

$$(k=1,2,\cdots m)\quad 0\leqslant r_{ij}\leqslant 1$$

计算完相似性系数后,就得到了相似性系数矩阵(或模糊矩阵)$\mathbf{R}$,即 $\mathbf{R}=(r_{ij})_{n\times n}$。这样得到的 $\mathbf{R}$ 具有对称性,然后再改造 $\mathbf{R}$,使之成为模糊等价关系 $\mathbf{R}*$,改造可用传递闭包法,即计算 $\mathbf{R}\cdot\mathbf{R}=\mathbf{R}_2$, $\mathbf{R}_2\cdot\mathbf{R}_2=\mathbf{R}_4$, … 直到使 $\mathbf{R}_k\cdot\mathbf{R}_k=\mathbf{R}_{2k}=\mathbf{R}_k$ 为止。$\mathbf{R}*=\mathbf{R}_k$ 聚类时用不同的值来截 $\mathbf{R}*$,将得到不同的聚类结果。

何道生对性状加权后,采用模糊数学对 356 株参考菌株和临床分离菌株进行了数值分类。

3.3 利用计算机软件进行数值分类

系统聚类法及模糊聚类法,编程需要利用计算机完成。目前也有专用软件(如 MINTS),模糊数学软件。对于各菌株之间的聚类问题也可利用国际流行软件如 SAS、SPSS、BMDP 及 STAT 等。其中 SAS 软件提供了 11 种不同的样品聚类形式,即:平均法(AVERAGE),重心法(CENTROID),最长距离法(COMPETE),非参数概率密度估计法(DENSITY),最大似然法(EML),FLKEXIBLE 法,MCQUITTY 相似分析法,中位数法(MEDIA),最短距离法(SINGLE)两阶段密度法(TWOSTQAGE),最小方差法(WARD)等。

3.4 其它方法

朱厚础利用计算机进行数值分类做了许多工作。在论文中[29]提到了以下三种数值分类法:主成份分析,主坐标分析,非线性映象法等。

细菌数值分类是一个客观、准确、快速的方法,已得到广大分类学者普遍关注并予以实践。但是,数值分类作为微生物界的一种新的分类方法,由于其短暂的发展历史,使得这种分类方法尚存在着某些难以克服的缺点。另外,数学上分类的理论工具较完善,但在细菌分类方面有差距。比如,在各种数值分类方法中,相似性系数时诸多计算公式有哪些适合于细菌分类?有待于同仁们在实践中不断修定、总结、充实,以逐步完善数值分类这一学科,使之更好地服务于细菌分类。

(下转第 154 页)

(上接第 190 页)

参 考 文 献

[1] Sneath P, Sokal R.数值分类学：数值分类的原理和应用. 第一版, 北京：科学出版社, 1984.

[2] 林万明. 细菌分子遗传学分类鉴定法. 第一版, 上海：上海科学技术出版社, 1990.

[3] 周方, 陈宁庆. 第四届分析微生物学学术讨论会论文集, 中国微生物学会, 1～15.

[4] Sneath P H A. Numerical Taxonomy, In Bergey's Manual of Systematic Bacterilogy, ed by Frieg. N. and J. G. Holt, London, 1984, 15～17.

[5] Jilly B J. Int J Bioneal Compot, 1988, **22**: 107～119.

[6] 徐浩. 微生物学通报, 1974, 3: 32～35.

[7] 徐浩, 江慧修, 乔宝义, 等. 微生物学通报, 1975, **15**(1): 31～36.

[8] 马俊才, 赵玉峰, 王大翔, 等. 微生物学通报, 1990, **17**(4): 210～214.

[9] 马俊才, 赵玉峰, 王大翔, 等. 微生物学通报, 1992, **19**(4): 193～196.

[10] 何道生, 余满松, 毛运平, 等. 微生物学通报, 1990, **17**(4): 214～217.

[11] 李虹, 刘明富. 微生物学通报, 1991, **18**(1): 34～37.

[12] feltham R K A, Wood P A, Sneath P H A. Journal of Applied Bacteriology, 1984, **57**: 279～290.

[13] Shoshana B, Lapage S P, Curits, H A. *et al*. Journal of General Microbiology, 1984, **57**: 279～290.

[14] 冯日孝. 微生物学通报, 1985, **12**(3): 111～114.

[15] Barry H, Christine A D, Claire A P. Fermentative Bacteria Journal of General Microbiology, 1986, **132**: 3113～3135.

[16] Willcox W R, Lapage S P, Shoshana B, *et al*. Journal of General Microbiology, 1973, **77**: 317～330.

[17] 李钦. 微生物学通报, 1979, **6**(1): 25～30.

[18] 陈文新, 骆传好. 微生物学通报, 1986, **13**(3): 133～141.

[19] 李多川, 洗崇尧. 微生物学通报, 1993, **20**(6): 323～327.

(下转第 160 页)