实验技术 微生物数值分类法简介

徐 浩

（中国科学院微生物研究所，北京）

微生物的分类归群工作是微生物的利用与防治工作中必不可少的工作。微生物——例如细菌——的分类，过去都是按照传统的分类法进行的。但近十多年来国际上又出现了一种不同于传统方法的新的分类法，这就是数值分类法，过去也有人称为数码分类法（Numerical taxonomy）。

从原理上看数值分类划群的标准与传统的方法有两种不同之处，即：

（1）所取的各项特征在数值分类法中一般是依同等重要性而衡量的，也就是说各种特征在定归群时都有相等的份量。

（2）归群时通常是以单连锁（即两两个成对地）比较的方式求相似值而归成簇群（Cluster）。

数值分类法在实验资料的整理上一般是依靠电子计算机，而不是象传统方法那样要靠人的手工式的劳动。而且它整理出的结果不以传统的定出学名的方式，归到属或种等等，而是整理成表观群（phenon），它往往是以图或表的方法表示之。但是迄今为止用这种非常不同的方法得出的结果，却往往是和传统分类方法相一致，看来这种方法的可信性是毋庸置疑的。但数值分类作为一种方法或思路还正在试验中，开拓中，改进中。它是传统方法的一种，目前还不是传统方法的代替者。

数值分类所根据的同等重要性的原则远在1757年就由 Michel Adanson 在进行软体动物分类时提出了，因此在国外文献中对等重要性原理又叫做阿德逊氏原理（Adansonian principle）[1]。在1956年 Sneath 首先把它用在细菌分类中[2]。由于实验资料近年来可以用电子计算机整理，所以大为减轻了分类工作者的脑力劳动量，而且使得工作程序更为自动化和客观化，从而减少了许多争议，也减少了在资料整理中必须大量参考前人文献的麻烦。

但另一方面在使用数值法时，同等重要性原则和单连锁群比较法却也引起不少疑虑。这是因为有人认为同等重要性原则没有分别主次，从而未能突出主要矛盾，而单连锁群则未能表示出同一级别的分类单位的共同特点。但是这些缺点并不是不可克服的，例如可以将某些特征“加权衡重”，即对它乘以大于1的系数，以修改等重衡量的原则；而对诸单连锁群也可以用计算机找出其共同的特征，使诸单连锁群之间建立起连系来，以找出共同特点。在实际工作中传统分类和数值分类也并不是截然在方法上互不相容的。例如当我们进行细菌的数值分类时，我们实际上就是承认原核的重要性是超越于一切其它特征之上的。而在传统分类中当某些株系的特征不能完全符合检索表的条款时，有时并不妨碍这些株系的依照大体趋势所定的分类位置，这在原理上看来是违反了多连锁群的传统办法的，从而实际上是承认了单连锁群的存在。

下面谈一下数值分类法的原理和资料整理办法。

（一）什么是同等重要性原则，为什么要采用这种原则

在进行分类时，分类者提出一系列问题，并从分类对象上找出答案。找答案的办法是通过观察或实验，然后将所得到的答案用种种方法进行整理。例如：细胞是什么形状 $\xrightarrow{\text{观察}}$ 圆形或杆状（答案）；机体能否运动 $\xrightarrow{\text{相差显微镜观察}}$ 动或不动（答案）。将这些结果通过一定的方式整理后，将每个处理分类单位（Operational taxonomic unit «OTU»）*，例如说将某个菌株，定到某个分类单位（taxon）中去[3,5]。

在向分类对象提问题及取得答案时，往往是相同的，但同样的答案，整理的结果差别却常常很大。我们可以看一下下面的例子在传统分类整理中可能发生的一些情况。这种差别往往是由于人们对特征的重要性的程度没有建立同样的认识或标准而引起的。

（1）传统分类的整理法：往往是依照资料整理者对诸特征的重要性的认识分成不同重要性的层次，再依典型化的二分歧法进行整理。今以具两对特征的四个细菌菌株说明之：

（a）细胞是圆的，能运动的。

（b）细胞是圆的，不能运动的。

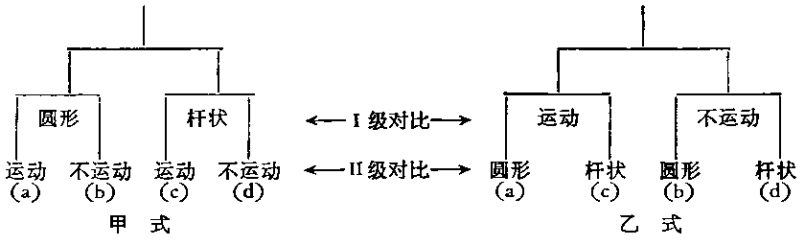
（c）细胞是杆状，能运动的。

（d）细胞是杆状，不能运动的。

这样随着个人对特征重要性的主观认识，就可以有下列两种分法。

依照甲式整理的结果，则（a）和（b）之间的关系较之（b）和（d）之间的关系更为密切。反之依照乙式整

* 也有人称为“操作分类单位”。



甲 式　　乙 式

理的结果则(a)和(c)较之(c)和(d)更为接近。

如果考虑到分类的目的之一是要求揭示自然的亲缘关系，那么在上述两式中到底哪一种更客观地反映了现实呢？在这种类似情况中是存在着争论的，会有许多种不同的观点，颇难定论。但是造成分歧的原因却很清楚地是由于在整理时没有将各种答案作等重衡量，有人更重视运动的重要性，有人更为重视形态的重要性等等。

在实际分类工作中，还有另外一种造成分歧而且使得各家说法难以进行比较的是对同一级的分类单位，例如在一个种中，在划分时有人作过 n 级对比，而另外一个人则可能要做 $n+a$ 级对比。也就是说有人可能比过 5 次，有人则比过 7 次。这样就使得同样的一个种的概念，在实际情况下，概括范围却悬殊颇大，我们常常称之为大种化或小种化。到底应当按照哪一类划分法，一直是由分类工作者依据他自己的经验，甚至其喜好而定，缺乏统一的，客观的标准。为了补救上述缺陷，可以采用阿德逊氏的等重衡量法。

(2) 等重衡量法：人们可以选取在细胞形态上，细胞排列上，菌落上，液体培养基中，生理上，生化上，营养要求上，对抗生素敏感性上等诸方面，大致在数量方面均匀配置的 40—60 个"问题"，通过观察或实验在工作对象上寻求"答案"。然后将这些答案不予歧视地、视为等重地进行比较，再求诸 OTU 之间的相似性。为什么要选用数目如此多的特征，这是因为所选用的特征必须是彼此独立的，例如鞭毛和运动虽然是两项特征，但却是一对密切相关的特征，因此只能选取其中的任一种。在我们选取的问题中，有许多其内在关系尚未能搞清，因而只能靠我们多选一些以予以补救。一般言之答案的可致信程度正比例于测出的数目的平方根。例如假设我们采取了 100 项答案和 25 项答案的两种实验结果，则它们的可致信程度的比例就是：

$$\frac{\sqrt{25}}{\sqrt{100}}=\frac{5}{10}=\frac{1}{2}$$

因此取 100 项答案时比取 25 项时可致信程度要高出一倍。因此在人力物力允许的条件下要尽可能地多做一些观察或实验项目。

(二) 资料的整理

在决定了特征的种类及数目，得到资料或数据后，紧接着就是进行整理，以便分析结果。在数值分类法中就是将各 OTU 所得的资料逐对地比较，以求得两 OTU 单连锁比较的相似值。逐对比较时可以用纸条法，或穿孔卡片法，但是这样比较时所耗费的人力极其巨大，很少有实用价值，现在一般是用电子计算机去代替人力去进行整理和比较。但不论用人力或机器其原理是一样的。下面用一个简化的例子说明资料的整理步骤。

例：今有 4 个菌株，每个菌株作了 4 个特征。其整理步骤为：

首先将实验结果列成表

表 1

菌株 \ 特征	1	2	3	4
A	+	+	−	−
B	−	+	+	*NC*
C	+	+	−	+
D	+	+	+	+

其中 *NC* 是缺项(英文原文是"不计"«not considered»)。当列出这个表格后，第二步就是将各 OTU 进行比较。四个菌株两两作单连锁比较，依下表可以看出需要比较的次数。

	A	B	C	D
A	AA	AB	AC	AD
B	BA	BB	BC	BD
C	CA	CB	CC	CD
D	DA	DB	DC	DD

在上述矩阵中，除了直线穿过的地方是每个菌株自己比自己，自然是 100% 的相似，从而无意义以外，直线的左下方和右上方是相同的，因此实际真正进行比较的部分只是一个矩阵的三角形部分，或可称之为三角矩阵。在 m 个菌株，每个 m 个特征时，它是 $m\times m$，再减去 m 个，然后被 2 除，即：

$$\frac{m^2-m}{2}=\frac{m(m-1)}{2}$$

实际上也就是在 m 个中，每次取 2 个的组合。

$$C_m^n = \frac{m(m-1)(m-2)\cdots(m-n+1)}{n!},$$

即矩阵每边是 m 个，因此可比较 $m \times m = m^2$ 次，但其中有 m 个是自己比自己无意义要减去，同时又只能取三角形内的不重复的部分，因此是：

$$\frac{m(m-1)}{2}$$

若我们有 200 个菌株的资料要比较，则进行单连锁比较求相似值时要计算

$$\frac{200 \times 199}{2} = 19900 \text{ 次}$$

这只是计算的次数。由于在计算前更烦琐的是每株菌的 40—60 个甚至更多的特征要逐一核对是否相符，因此要核对至少 $19900 \times 40 = 796000$ 次。作 80 个特征甚至要核对将近一百六十万次，显然这项工作不但异常烦琐，而且是一般情况下人力不堪负担的工作量。因此使脑力劳动机械化是必不可少的。

在进行比较时，也就是例如依据表 1，求每一对菌株间的相似百分比。它至少有两种方法。即计入或不计入同为负号的情况，或计不计算负相符。例如两株菌都不能运动，算不算具有相同的特征呢？依据计不计入负相符可有两种不同的相似值计算公式，即：

$$S_{(I)} = \frac{\Sigma +}{\text{全体} - [(\Sigma -) + (\Sigma nc)]} \qquad \text{(a)}$$

在这种 $S_{(I)}$ 求相似值的公式中，只计入正相符。而另外一种计算公式是：

$$S_{(II)} = \frac{(\Sigma +) + (\Sigma -)}{\text{全体} - \Sigma nc} \qquad \text{(b)}$$

在这种 $S_{(II)}$ 求相似值的公式中，正相符和负相符全都计入在内。不论那种计算法 nc 均被从全体数中减去，可知 nc 与负相符不同，它使分母值发生改变。在实践中 nc 往往是缺项，实验中缺项过多会导致 S 值上升，但同时会使实验的可致信度下降。

在计算出 S 值以后，就要进行所谓“簇群分析”(Cluster analysis)。“簇群分析”得名的由来是因为各个 OTU 是两两比较得到的集合物。因此先要找出一对最高水平相似值的 OTU，然后再围绕它们找出与之比较过，而且相似值在同一水平上的一些 OTU，如此组成“簇群”。兹将表 1 的资料用 $S_{(I)}$ 计算其 S 值如下：

$$S_{B-A} = \frac{1}{4-1} = \frac{1}{3} = 0.33$$

$$S_{C-A} = \frac{2}{4-1} = \frac{2}{3} = 0.66$$

$$S_{D-A} = \frac{2}{4} = \frac{1}{2} = 0.50$$

$$S_{C-B} = \frac{1}{4-1} = \frac{1}{3} = 0.33$$

$$S_{D-B} = \frac{2}{4-1} = \frac{2}{3} = 0.66$$

$$S_{D-C} = \frac{3}{4} = 0.75$$

当算出 S 值后就可以整理成下列的各种表示图表。

表 2　机体簇群表

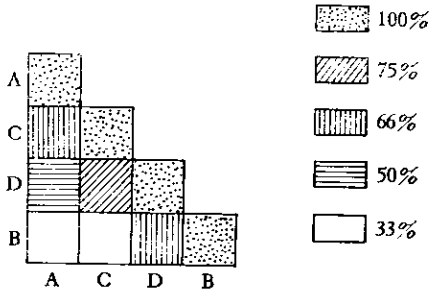
% S	簇群
75	*C–D*
66	*B–D*, *C–A*
50	*D–A*, *C–B*
33	*C–D*, *A–B*

对上述簇群整理方法要作一些说明。即在相似值 75% 的簇群只有 *C–D*；而在 66%时，则有 *B–D*, *A–C*；在 50% 时，原来只有 *D–A* 比较出过，但 *B–D*, *C–A* 都有着 66% 的相似性，因此要包括 *C*—*B*。对相似性 33% 的比较也是这样。这样就有四级水平的簇群，如表 2 所示。在这里就暴露了单连锁比较法的一项缺点，即 S_{C-B} 直接比较时明明是 33%，但这里却在 50% 中出现。这是因为在这种形式的簇群分析情况下，对任一 OTU 而言，与群内任一 OTU 达到所列的相似性水平，就足够使之放在这一群内，但它可能与这一簇群中的其它 OTU 只有很低的相似性[1,3]。关于这一问题的讨论，本文不能详及[3,4]。此外这样形成的簇群，并不一定具有共同的特征，它特称为多元群 (polythetic group)，它是与传统分类的单元群 (monothetic group) 是相对而言的。但由于所取的特征有限，通常是 40—60 个左右，各簇群中重复的特征必然会出现，因此在某种程度上多元群也反映出了单元群的特色。从而我们通常可将 $S = 75\%$ 时的诸 OTU 视为一个种，将 $S = 65\%$ 的诸 OTU 视为一个属[4,5]。

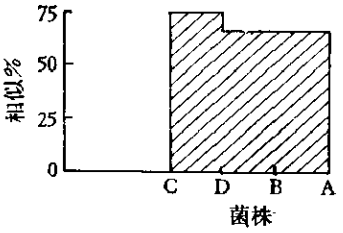
可以将表 2 整理成一个三角矩阵，即：

A	100			
B	33	100		
C	66	33	100	
D	50	66	75	100
	A	B	C	D

或用不同的阴影画成：

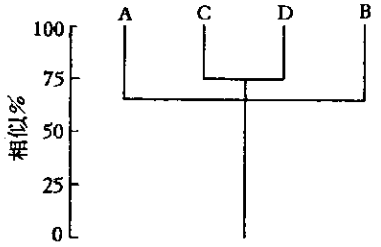


以上的资料还可以整理成“边线图解”(skyline diagram)



绘制这种图的方法是取座标纸,纵座标为 S 值,横座标为菌株。菌株次序可依整理过的机体簇群表,然后依菌株在该表中的最高 S 值定纵座标,横向连接而得。

此外还可以整理成枝叉图 (dendrogram),例如将表2绘制成:



这张图的绘法是把 C-D 在 75% 处连接,然后在 66%处将全部菌种连接了起来。请注意枝叉图不是系统树,这是一种并不一定能反映亲缘关系的方便直观的表示方法。

(三)特征的符号化问题

最后我们还要谈一下怎样将特征的记录符号化以便于整理的问题。

当我们记载微生物的特征时,我们会发现特征分成两大类,一种是二态特征,如能运动或否,可表示为+或−。另外一种则为多态特征即状态上有着程度或量的差别(如对某种抗菌素的敏感性),这时要把它分成以符号表示的诸级别,从而便于衡重及资料分析。常用的有下列二法:

I 法,记为:

符号 / 级别	w	x	y	z	衡重所得值
0	−	−	−	−	$\frac{0}{4}=0$
1	+	−	−	−	$\frac{1}{4}=0.25$
2	+	+	−	−	$\frac{2}{4}=0.5$
3	+	+	+	−	$\frac{3}{4}=0.75$
4	+	+	+	+	$\frac{4}{4}=1$

这种方法又叫递加法,因为分子的大小递加。这种记法往往会加大特征之间的差别。

另外一种记法叫做非递加法,即:

II 法,记为:

符号 / 级别	w	x	y	z	衡重所得值
0(不可查觉)	−	nc	nc	nc	$\frac{0}{1}=0$
1 级(弱)	+	+	−	−	$\frac{2}{4}=\frac{1}{2}=0.5$
2 级(中)	+	nc	+	−	$\frac{2}{3}=0.66$
3 级(强)	+	nc	nc	+	$\frac{2}{2}=1$

在记法 II 中,用下列公式求其衡重所得值:

$$\text{衡重所得值}=\frac{\Sigma+}{\Sigma\text{全体}-\Sigma nc}=\frac{\Sigma+}{4-\Sigma nc}$$

这种方法中 nc 影响了分母的值,这种记法往往会缩小对比特征之间的差别。

值得注意的是在多态特征的符号化表示法中,在某种程度上已经包含了非等重衡量的因素。这是因为我们将某一特征很强时才看做 1,其它均乘以一个小于 1 的系数。假使我们将某一特征全强时作为 4,则 I 法可记成:4,3,2,1。这就是乘以大于 1 的系数的加权衡重。

另外人们也可以在一些单连锁群的诸 OTU 的各个特征中,找出出现概率很高,例如接近于 1 的一些特征,从而真正客观地找出真正值得特予重视的特征。也可以找出诸簇群间的最低相似值,合并成多连锁群。这些都是可以尝试去做的。因此受到有些工作者非难的等重量原则和单连锁比较法,并不是数值分类法中的不可医治的痼疾,只要客观上需要,人们完全可以将它修订的。因此数值分类的特点是这种方法使脑力劳动机械化了,从而使得分类工作者可以节约下有用的精力去考虑更为重要的问题。

参考资料

[1] Lockhart, W. R. and J. Liston: Methods for Numerical Taxonomy. 1970.

[2] Sneath, P. H. A.: *J. Gen. Microbial.* 15, 70. 1956.

[3] Ainsworth, G. C. and P. H. A. Sneath: Microbial Classification, pp. 289—332. 1962.

[4] Sokal, P. R. and P. H. A. Sneath: Principles of Numerical Taxonomy. 1963.

[5] Sherman, V. B. D.: 细菌属的检索(第2版,第一、二章译稿)。1967。