技术与方法

根瘤菌数值分类方法探讨

陈文新　骆传好
（北京农业大学微生物专业）

数值分类是指用数值分析方法，借助电子计算机将运筹分类单位（OTU）（如细菌菌株）按其性状状态的相似程度归类成表观群（phenon）。为了揭示生物分类单位间的真实关系，需要尽可能多的特性测定，将这些信息转换为数值，对这些数值进行运筹处理，其计算量往往十分巨大，必须依靠电子计算机来完成，所以直至本世纪四十年代电子计算机技术兴起，才为数值分类学发展创造了条件，使它成为生物分类学在本世纪中一项重大突破，把一门描述性，定性的分类学提高到精确的定量水平上。

数值分类的一个主要原则是给分类单位的各性状状态在建立分类单元（简称类元）时以相等的地位，即"等重"原则。这原则是林奈的同时代人植物学家 Michel Adanson 提出的，所以数值分类法又叫 Adanson 法。数值分类法引入微生物研究领域是英国细菌学家 Sneath 五十年代开创的。六十年代他又建立了许多分类分析方法，为微生物数值分类奠定了基础。继 Sneath 之后，日本细菌学家塚村（Thukamura）对分枝杆菌（*Mycobacteria*）进行了数值分类，日本的林江沢对球菌科（Coccaceae）分类进行定量研究，并提出"中心种"的概念。波兰的微生物学家 Kurylowicz 做了链霉菌属（*Streptomyces*）的数值分类研究。这些工作都为微生物数值分类研究起了很好的推动作用。

数值分类由于它根据尽可能多的性状状态所反映的信息，借助于数学方法和计算机处理，减少了工作者的主观偏见，所以，数值分类在分类关系的估价和类元的建立上都是客观的，明确的和可重复的[1]；它使分类过程自动化，大大提高了工作效率；还可进行数据存贮，自动编制目录、检索等；它的定量分析对分类单位提出了更大的鉴别差异，在类元划分上更灵敏，故可得到更好的分类系统和检索。因此数值分类在七十年代在国际上得到迅速发展。笔者从 1981 年开始，结合对我国根瘤菌资源调查、开发利用，从事根瘤菌数值分类研究，先后三次共用了 150 个菌株，测得 200 个左右的编码性状，进行根瘤菌分类，得出了与国际上相一致的关于根瘤菌的分类体系。也发现了我国土著根瘤菌新的表观群，与此同时，我们对选用不同性状状态集、不同相似性系数、不同聚类方法进行了比较、对聚类分析在根瘤菌分类上的应用作了方法学上的探讨，本文是结合我们的工作对数值分类方法要点进行介绍。

数值分类方法的一般程序是：分类单位及单位性状的选择 ⟶ 性状的测定及收集 ⟶ 性状的数据编码 ⟶ 计算相似性系数值 ⟶ 聚类运算 ⟶ 分类结果表示 ⟶ 菌株鉴定。

（一）运筹分类单位（OTU）及单位性状的选取

运筹分类单位的选用应以工作目的为依据，如 Graham[2] 曾想探讨根瘤菌和与之相近的曾被分类在同一个科中的其它属的菌的关系，确定根瘤菌的分类地位。我们的目的是根瘤菌分类方法学探讨，因此只局限在根瘤菌菌株，且只注意了根瘤菌寄主的多样性及地理分布的多样性，包括老系统中的不同种，并引进了老分类系统中的模式菌株及与根瘤菌关系密切的土壤杆菌的菌株作参比。

单位性状是用于分类研究的基本信息单位，它是决定分类结果的关键，每一性状状态都应提供一项新的信息。

究竟需取多少性状，选择哪些性状，这些均为方法本身尚未确定的问题。Sneath[4] 曾经提

过需50个性状，后来他又提出希望有几百个性状。

性状选取的原则是应尽可能广泛而均匀地遍布于所研究的机体上，我们采用的根瘤菌性状包括形态的、生理生化的、营养的(碳源、氮源，维生素)、生长温度范围、生长pH范围、抗逆性(耐盐、耐酸碱程度、耐高、低温度、耐抗生素、染料、农药)等。

有些性状是不可采纳的：1. 无意义的性状；2. 逻辑相关性状，如运动性与鞭毛不能同时采用；3. 全同性状：在全部OTU上都相同的性状，如根瘤菌均是G^-菌。有些性状必须在测定后方得知其全同与否，那么需待测定后才能决定取舍。

不同的性状集在分类结果上占的地位如何，我们曾作过一点比较；将根瘤菌的营养性状集、对抗生素的敏感性状集及对染料的敏感性状集等的单项相似性树状图分别与全相似性的树状图作比较。其分群结果大体一致，即在一定相似性水平上快生型根瘤菌和慢生型根瘤菌都明显地归为两大类群，在较高相似性水平上各群的划分和全相似性值分析的基本趋势一致，以营养性状集的分群结果与全部性状分群结果更相一致。用不同性状集的相似性分类结果与全相似性结果所不同的是：菌群与菌群之间的密切程度有所变化，少数菌株的归属有所不同。比较来看，染料性状集的分类结果不如其它的理想，多为阶梯式的树状图，使划分类群的水平很难确定。染料性状用于分类的报道不很多，对染料反应的性状不宜多用。

(二) 性状编码

把实验中测得的性状状态的记录结果转化为计算机所能识别、运算的符号即为性状编码。这是一个较为复杂的问题，只有处理得好，才能得出好的分类结果；因为采用的性状有多种不同的状态，所以应采用相应的不同编码方法。

1. 两态性状：这是在微生物分类中采用最多的性状。如菌株对某种碳源利用与否，或某种酶的有无，其编码为：阳性结果用"+"表示，阴性结果用"−"表示（输入计算机时分别用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)

采用的性状中，有些包括有多种状态，即彼此两极端间有中间过渡状态，有定性定量之分，有序无序之别。

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，根瘤菌在加有不同浓度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，将一个定量多态性状转换成多个二态性状，谓之"加权"，递增编码如表1所示。

从表1看，随着根瘤菌耐盐能力的增加，编码性状为1的个数也作相应递增。Sneath倾向于推荐这种方法，我们对定量多态性状也采此法。

表1 递增编码法

OTUs	性状编码 / 性状 / 性状状态	NaCl(1%)	NaCl(2%)	NaCl(3%)
		1	2	3
1	在1% NaCl上不生长	0	0	0
2	在1% NaCl上生长	1	0	0
3	在2% NaCl上生长	1	1	0
4	在3% NaCl上生长	1	1	1

3. 定性多态性状或称无序多态性状，性状的多种状态无顺序可言，最难准确编码。一般也可将多态转换成多种二态性状，如表2。

对石蕊牛奶反应则采用了表3所示的方法。

表中采用了NC符号。酸凝是产酸量大的一种表现(从这一意义上讲就有递增的成分)，2

表2 加权非递增编码

OTUs	菌落颜色	二态性状		
		1	2	3
1	红	1	0	0
2	黄	0	1	0
3	蓝	0	0	1

号菌不产酸，也就谈不上酸凝；同样，3号菌由于生长快，一开始就还原了石蕊，使之失去了酸碱指示的作用，所以也用了NC符号。

4.有层次的多态性状，有的性状有初级次级之分，如鞭毛有无是初级性状，而鞭毛的位置是次级性状，编码方法可如表4。

性状编码完成后，都排列成顺序号，形成一

表3 加权递增与非递增混合编码

性状状态 \ OTUs		1	2	3
石蕊牛奶反应	产酸	1	0	NC
	酸凝	1	NC	NC
	产碱	NC	1	NC
	胨化	1	1	1
	还原	0	0	1

表4 初级、次级性状编码

OTUs	编码 \ 性状层次 / 性状状态	初级性状	次级性状	
		鞭毛有无	鞭毛极生	鞭毛周生
1	无鞭毛	0	0	0
2	鞭毛极生	1	1	0
3	鞭毛周生	1	0	1

表5 原始特征编码表

编码 / 特征 / 单位性状 \ 菌株号	1	2	3…31…55
1.d-葡萄糖酸钠的利用	0	0	1…1…1
2.D(+)纤维二糖的利用	1	1	1…1…0
⋮			⋮
78.对1% NaCl的耐受性	1	1	1…1…0
79.对1.5% NaCl的耐受性	1	1	1…1…0
80.对2% NaCl的耐受性	1	1	1…1…0
⋮			⋮
208.氧化酶活性的有无	1	1	1…1…1

表内"……"为删节号。

个性状(原始数据)矩阵，即可输入计算机。如表5。

（三）相似性系数的计算

相似性系数是用来描述被比较的OTU对偶间相似程度的量值。又分Q技术及R技术，Q技术是指全部性状作数值分析对OTUs进行聚类，是生物数值分类常用的以分析OTU间的分类关系；R技术则相反，是通过OTUs对性状进行分析，找出性状之间的相互关系，生态研究常采用。

相似性系数种类繁多，算法各异，常用的不下几十种，大体上分为相似性和相异性系数两类。前者测度相似性，后者主要是各种距离系数，测度差异，两者互逆，可以一定方法换算。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，适于二态性状数据。为了说明各公式的含义，先介绍各公式中所列a，b，c，d所表示的量，见表6。

Austin等[3]曾用141株肠杆菌科的菌株，用了240个性状编码，用36种系数计算相似性，用平均连锁法进行聚类，他们的结果表明有

表6 字母含义

OTUs		OTUs	
		1	0
OTUs	1	a	b
	0	c	d

a 表示两个 OTU 的性状编码皆为 1 的个数，称正匹配。
b. 表示一个 OTU 为 1，另一个 OTU 为 0 的性状个数，为错配。
c. 表示一个 OTU 为 0，另一个 OTU 为 1 的性状个数，亦为错配。
d. 表示二个 OTU 皆为 0 的性状个数。称为负匹配。

15 种系数可用于细菌分类，其中

$$SH\left(\frac{a+d-b-c}{a+b+c+d}\right)$$

和

$$S_{TD}\left(\frac{b+c}{a+b+c+d}\right)$$

与

$$S_{SM}\left(\frac{a+d}{a+b+c+d}\right)$$

所得的结果没什么不同，S_0 及 S_0(公式略)所得结果与 S_{SM} 相近，$S_D\left(\frac{2a}{a+b+c+d}\right)$ 明显节省计算机时。

我们曾用 S_{SM} 及 S_D 和 S_P 三种公式计算供试根瘤菌所得相似性值，用平均连锁法聚类的结果进行了针对这三种系数的比较（在 S_{SM} 式中把比较的负匹配 d 也看做是两个 OTU 间的相似状态；在 S_D 式中是加倍了正匹配，用 2a，而不计负匹配 d，$S_P = 1 - D_P$，D_P 是通过 D_V 矩阵求得的，容后介绍）。三种公式的聚类结果大体一致，S_{SM} 和 S_P 的结果更接近，但 S_P 所得的结果中土壤杆菌群不如 S_{SM} 得的结果一样与根瘤菌在属的水平上分开，用 S_D 公式则更差一些，它将土壤杆菌与苜蓿根瘤菌混在一个类群里，说明 S_D 的敏感性差。所以我们认定进行根瘤菌分类还是用 S_{SM} 相似性系数为好。

由相似性系数得到 t × t，OTU 间相似性矩阵(略)。

(四) 系统聚类(或等级聚类)

根据相似性系数值 (S) 或相异性系数 (即 $D_T = 1 - S$) 对 OTU 进行系统聚类归群。通俗地讲是将类由多变少，最后归成一个。在聚类时，我们设测验的 n 个性状为 n 维空间，供试的七个 OTU 为 n 维空间中的 t 个点。系统聚类的一般步骤为：1. 样品(即 OTU)各自成一类；2. 计算各样品之间的距离(或相似系数)，将最近的两个样品并成一类；3. 计算新类与其余各类的距离，再将距离最近的两类合并，如此类推，直到所有样品归为一类才停止。

因类与类之间用不同的方法规定距离，就产生了不同的系统聚类方法，最常用的方法有四种，即单连锁法，全连锁法、平均连锁法及可变连锁法。

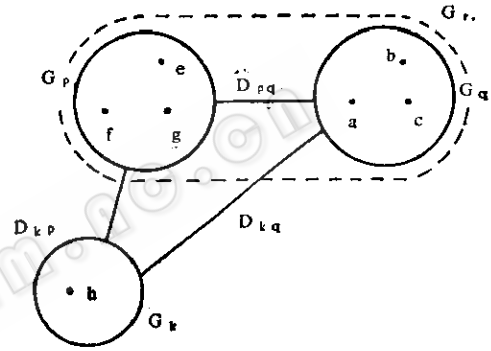

图 1 聚类法图示

图 1 表示四种方法计算类与类之间的距离各不相同，字母 G 代表类，D 代表距离，图上已有 k、p、q、三类，每类中已有各个 OTU，如：a、b、c、d、e、f、g 和 h。

单连锁法计算的是图 1 中 p 群(G_p)和 q 群(G_q)中两个最近的点 g、a 间的距离；全连锁法计算的是 G_p 和 G_q 中两个最远的点 f、b 间的距离；平均连锁法，也称不加权算术平均法。它计算的是两群中各 OTU 间的平均距离，如：$\overline{la} + \overline{lb} + \overline{lc} + \overline{fa} + \overline{fb} + \overline{fc} + \overline{ga} + \overline{gb} + \overline{gc}$；可变连锁没有明显的几何学意义。

Wishart 于 1969 年发现了一个系统聚类八种方法递推公式的统一形式，这为编制计算机统一程序提供了极大的方便，统一公式为：

$$D^2_{Kr} = \alpha_p D^2_{Kp} + \alpha_q D^2_{Kq} + \beta D^2_{pq} + r|D^2_{Kp} - D^2_{Kq}|$$

(五) 聚类结果表示

有很多表示方法，一个是相似性矩阵，在矩

阵中 OTUs 按聚类方法所给的顺序重新排列，然后矩阵的小方格涂成暗色，相似性最高的用最黑的色调。在这些涂黑的 S 矩阵中，聚类用黑三角表示(图略)。较常用的聚类分析结果是画出树状图(图 3)，在图中最紧密的小分枝代表最相似的 OTUs 类聚。

(六) 聚类结果检验

我们曾用 55 株菌，测定 208 个编码性状，采 S_{SM} 系数，用这 4 种聚类方法得出四个树状图表示的聚类结果。然后再采用协表相相关(Cophenetic Correlation) 系数(R_{CS})和分类学距离对四种聚类方法的优劣进行了评价以及结果检验。

所谓协表相相关是对既定树状图所包含的任何 OTU 对偶间的相似性得一协表相值，并为任一 OTU 集得出一个这样值的矩阵，如图 2。

OTUs	a	b	c	d	e
a	X				
b	1.0	X			
c	3.8	4.0	X		
d	4.4	4.2	2.6	X	
e	5.1	5.0	5.3	5.4	X

相似性矩阵 S

OTUs	a	b	c	d	e
a	X				
b	1.0	X			
c	4.1	4.1	X		
d	4.1	4.1	2.6	X	
e	5.2	5.2	5.2	5.2	X

协表相矩阵 C

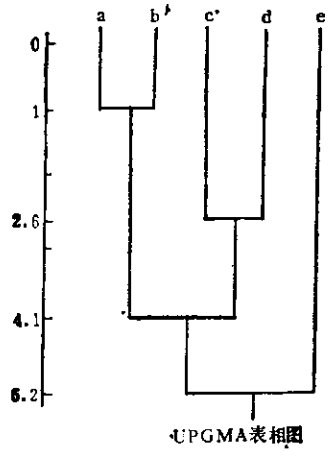

图 2 UPGMA 表相图

矩阵 S 含 OTU 间原始相似值，从 S 推导出来的树状图得到协表相距离制定矩阵 C，协表相相关系数 R_{CS} 是表示来自 C 和 S 矩阵的相应对偶间的相关，可由下式得出：

$$R_{cs}=\frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1}s_{ij}c_{ij}}{\sqrt{\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1}s_{ij}^{2}\right)\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1}c_{ij}^{2}\right)}}$$

i、j 分别表示不同的OTU，n 代表总的OTU数$\left(\text{这里用}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\text{即将对角线上的数值省去}\right)$。

通过比较 C 矩阵和原始相似性矩阵 S 得出各个协表相相关系数：

$R_{cs单连锁}=0.9807$　$R_{cs全连锁}=0.9670$

$R_{cs平均连锁}=0.9845$　$R_{cs可变连锁}=0.9746(\beta=0)$

$R_{cs可变连锁}=0.945(\beta=-0.25)$

从以上数据可知：各个 R_{cs} 值均很大，但其中平均连锁的最大，说明平均连锁的聚类最能反映出原始相似性矩阵，也就是说平均连锁给出了最佳表示法。即得到与原始相似性矩阵最相近的聚类结果。几种方法所得的树状图大同小异，我们以平均连锁的结果(图 3)进行分析。平均连锁在 75% 相似水平上全部 OTU 分为五群；其分群结果与当今国际上根瘤菌分类系统基本一致，惟独新疆中慢生型群与快生型群相近，而与普通慢生型菌群较远，所以我们用分类学距离进行了衡量。所谓分类学距离 (taxonomic distance) 设用一个球 (sphare) 或圆(circle)表示一个类元 (taxon) 的范围 (dimension)，它的半径 $r=(100-y)/y$，y 代表类元内的平均相似值，而类元间的距离 $\alpha=(100-X)/X$，X 代表类元间的平均相似值。

从计算的结果绘制于图 4I，II，从图中可知：不管是按平均连锁还是按全连锁聚类，第 VI 群都与 II、III 群相近，而与群 V 较远，从几个方面检测以及根瘤菌与寄主的关系，我们肯定平均连锁 (UPGMA) 的聚类方法对根瘤菌分类目前算是最佳聚类方法。

(七) 活力(vigor)和型式 (pattern) 分析

对菌株进行生理、生化、营养需求等性状检

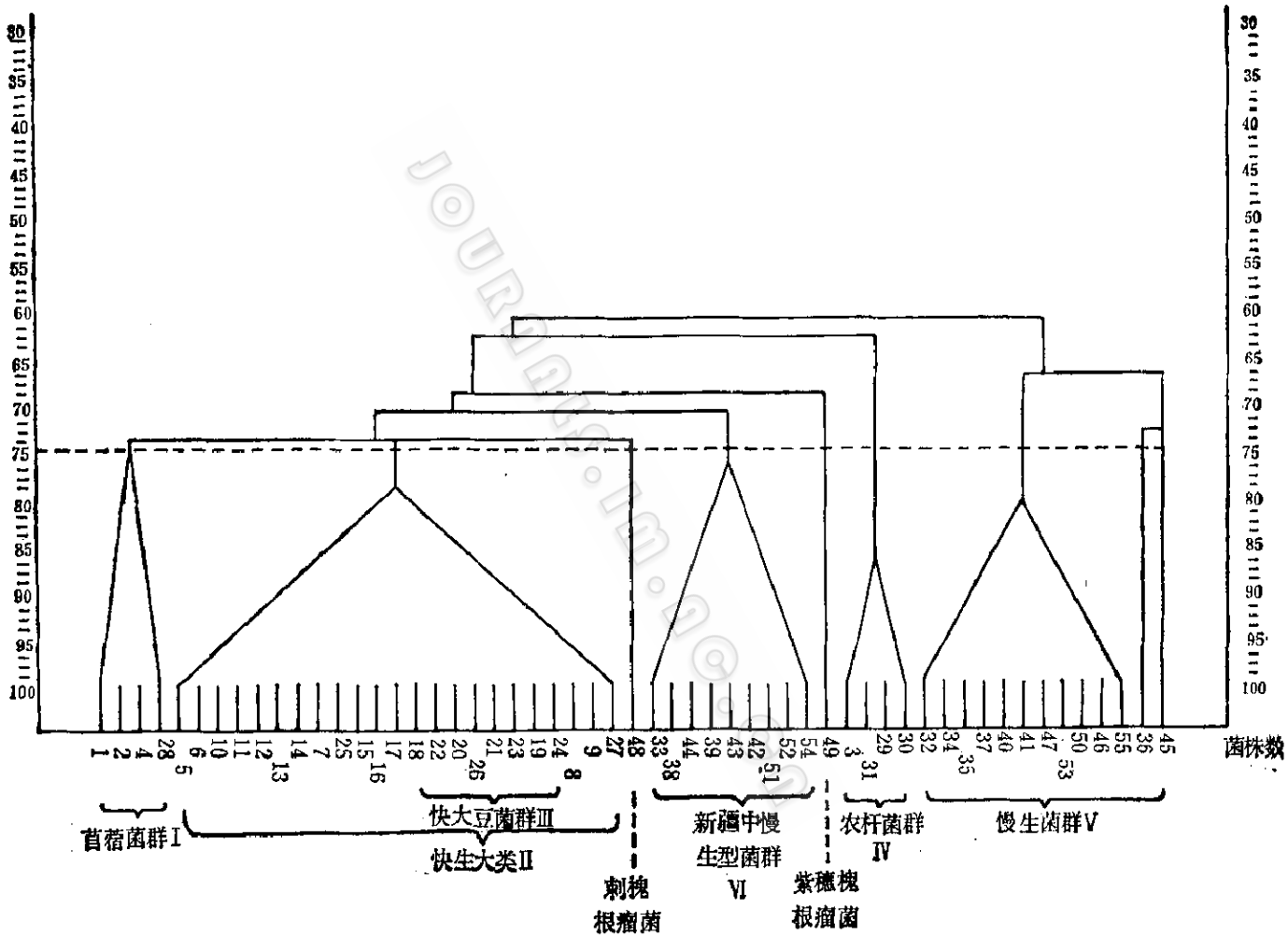

图 3　UPGMA 树状图

测时，常可能因测试菌株间的生长速度不一致而使结果受到影响，因为在试验中往往在同一时间观察结果，那么生长慢的菌株可能呈现弱或负反应，如果延长培养时间，则这些弱的或负的反应可能会转化成正结果，对这类情况应予以分析。Sneath 曾引进了活力差系数和型式差系数，他认为菌株间总的差异(D_T)由两部分组成，一部分是活力差(D_V)；另一部分是型式差(D_P)。总差异 $D_T = 1 - s$ (s 为相似值)即：$D_T = (b + c)/n$ (n 为测定性状总数)，Sneath 用正反应性状占全部测定性状的比率来表示活力，那么对 A 菌来说，其活力为 $(a + b)/n$，B 菌则是 $(a + c)/n$ A 与 B 的活力差

$$D_V = \frac{|b - c|}{n},$$

据 $D_T = \frac{b + c}{n}$，$D_V = \frac{|b - c|}{n}$ 可得：

$$\frac{(c - b)^2}{n^2} + \frac{4bc}{n^2} = \frac{(b + c)^2}{n^2}$$

设 $D_V^2 + D_P^2 = D_T^2$

于是 $D_P = 2\sqrt{bc}/n$，当 $D_V^2 = D_T^2$，即 $D_V = D_T$ 时 $D_P = 0$，那么，A、B 两菌之间的差异纯粹是活力差异，这很可能由生长速度造成，故在数值分析时对这类情况应给予分析。上面我们曾引用过 S_P 相似性系数 ($S_P = 1 - D_P$) 即排除了活力的影响。我们用 S_P 系数所得的结果与用 S_{SM} 的相近，说明活力差没有影响结果，也因为我们观察结果是根据根瘤菌生长快慢而在不同的时间进行的。

另一方面，可以认为 D_V 系数在一定程度

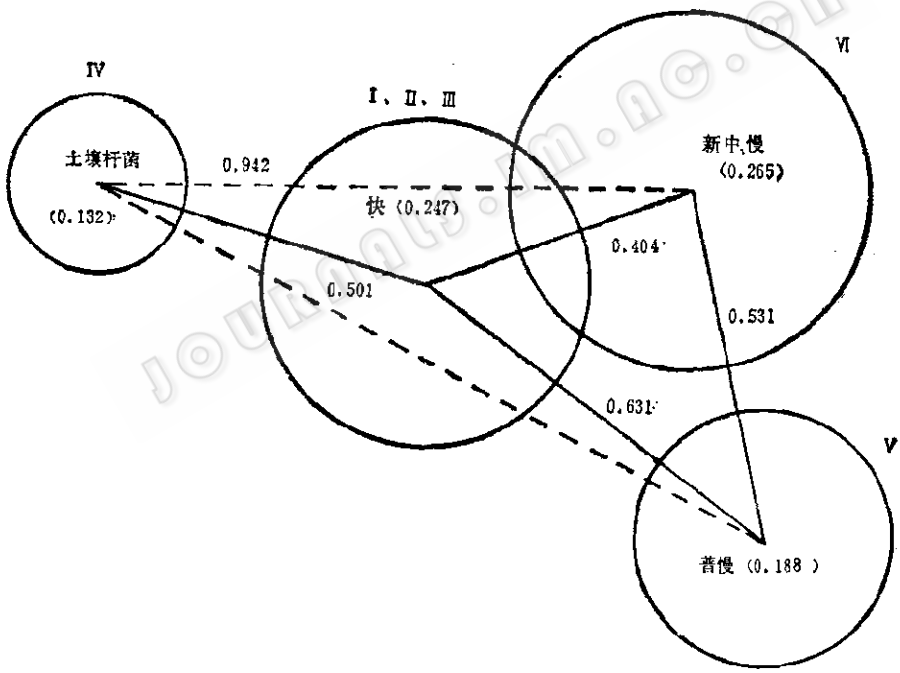

图 4-1 用分类学距离表示的类元间的相互关系(四大表观群的相互关系，按平均连锁法)

$$r = \frac{100 - \bar{S}\ 群内}{\bar{S}\ 群内} \quad d = \frac{100 - \bar{S}\ 群间}{\bar{S}\ 群间}$$

$\bar{S}$ 土壤杆菌=88.37%　　CV = 3.72%
$\bar{S}$ 快=79.75%　　CV = 25.40%
$\bar{S}$ 慢=84.15%　　CV = 3.94%
$\bar{S}$ 新中慢=79.02%　　CV = 6.63%
$\bar{S}$ 新中慢与土=51.49%　CV = 8.14%
$\bar{S}$ 新中慢与快=71.25%　CV = 8.10%
$\bar{S}$ 新中慢与慢=65.31%　CV = 7.97%

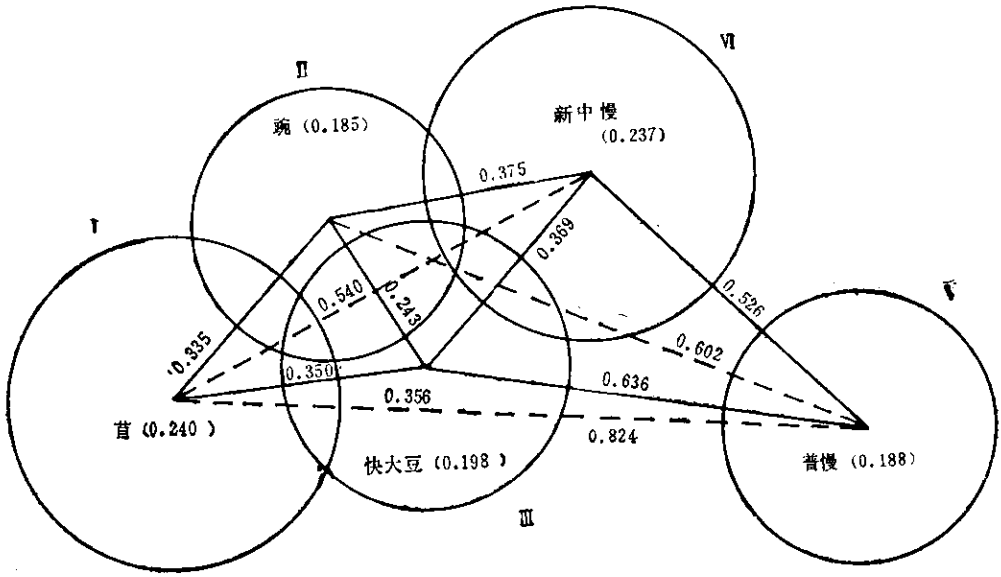

图 4-II 用分类学距离表示的类元间的相互关系（按全连锁法）

$\bar{S}$ 苜=80.67%	CV = 6.12%
$\bar{S}$ 豌=84.36%	CV = 5.46%
$\bar{S}$ 快大豆=83.49%	CV = 2.49%
$\bar{S}$ 新中慢=80.82%	CV = 5.12%
$\bar{S}$ 慢=84.15%	CV = 3.94%
$\bar{S}$ 新中慢与苜=64.93%	CV = 6.19%
$\bar{S}$ 新中慢与豌=72.75%	CV = 5.81%
$\bar{S}$ 新中慢与快大豆=73.06%	CV = 4.69%
$\bar{S}$ 新中慢与普慢=65.53%	CV = 7.90%

上体现出被比较的两个菌株生活力的强弱，因而，从选种的角度可以引用 D_V 系数找出生活力最强的菌株。

（八）中心株的确定

中心株是每类群的所有菌株中具有最大平均相似值的菌株，因此，中心株应该是典型的菌株，可以看作新种的模式标本。（图 3 中标有小箭头的菌株即是）我们曾将前一批根瘤菌数值分类各类群的中心株有意安排在后一批根瘤菌数值分类实验中，虽然这前后两次实验所用的菌株绝大多数不相同，性状编码也有所差异，但其结果在相应的四个大群中，有三群的中心株完全相同。另一大群的中心株不一致的原因是其中增加了一个亚群，故增大了该类群中的离散性。

（九）种群的鉴定

传统的方法是检索表的二歧排列，并倾向于分类单元之间用完全互斥的特征来加以相互区别。现在看来，用这样的方法作细菌鉴定并不理想，完全互斥的特征也很难找到。例如，一般认为：蛋白胨的利用、尼尔兰还原及 3-酮基乳糖产生可用作土壤杆菌与根瘤菌相互区别的特征，但对苜蓿根瘤菌来讲仍然区分不了。另外，鉴别特征过少，人为的因素，即实验的误差对结果的判断的影响会随之加大。故 Sneath 认为：宁可选取“那些重复性没有所希望的那样好的多项特征，而不用那些重复性相当好的少数特征作为鉴定标准”。

我们对鉴定特征的选取是从已分定的类群中选取出现频率大于 95% 的性状为鉴定特征，即群内 95% 以上的菌株具备（或不具备）的性状记为“+”（或“−”），而不足 95% 的菌株具备的性状记为“±”。为此，可根据类群的级别列出其鉴定特征。当电子计算机日益普及的今天，

可以将鉴定数据存贮在数据库，并可以随时增加新的信息和进行修订。分类学检索表也能存贮在计算机里，并能通过"人机对话"的形式而很方便地加以应用。

通过几年来的实践结果，我们认为数值分类的确是细菌分类的一个客观、准确、快速的方法；用于同质菌株的类聚，细菌种、属的建立，方法已臻完善。Sneath[4]在 Bergey's 系统细菌学手册第九版中将数值分类列为细菌分类方法之首，概述了以上的数值分类程序及要领，并肯定用相似性分析将菌株排列成的表元（phenetic groups）是广泛地等同于类元（taxa）的。他指出已普遍发现在约 80% 相似水平形成的表元相等于细菌的种。

参 考 文 献

[1] 赵铁桥译、P. 史尼斯等著：数值分类学-数值分类的原理和应用，科学出版社，北京，1984 年。

[2] Graham, P. H.: *J. Gen. Microbiol.*, **35**: 511—517, 1964.

[3] Austin, B. and R. R. Colwell: *Int. J. Syst. Bact.*, **27**(3): 204—210, 1977.

[4] Sneath, P. H. A.: Numerical Jaxonomy, In Bergey's Manual of Systimatic Bacteriology 9th, ed by Frieg, N. and J. G. Holt, Vol. 15—17, 1984.