微生物学报 27 (4): 306-317, 1987
Acta Microbiologica Sinica

系统聚类分析在细菌全细胞脂肪酸模式识别中的应用*

朱厚础
(军事医学科学院生物工程研究所，北京)
周方 唐光江
(军事医学科学院微生物流行病研究所，北京)

用欧氏距离系数和指数相关系数，结合8种常用的系统聚类算法，对用毛细管柱气相色谱法绘制的34株莫拉氏菌（*Moraxella*）及其类属菌和13株嗜肺军团杆菌（*Legionella pneumophila*）的全细胞脂肪酸气相色谱图，进行了聚类分析。比较了欧氏距离系数的8种系统聚类算法所得的聚类树状谱。结果表明，莫拉氏菌与嗜肺军团杆菌可以明确区分。在莫拉氏菌中，我国分离的两个新种与目前该属的主要标准株也能明确区分。两种相似系数中，欧氏距离系数的聚类结果较好；8种系统聚类算法中，最长距离法和类平均法的聚类结果较好。

关键词 系统聚类分析；细菌全细胞脂肪酸模式；莫拉氏菌；嗜肺军团杆菌

1963年 Abel 等[1]证明了利用全细胞脂肪酸气相色谱（GC）分析进行细菌分类鉴定的可行性。此后，Drücker 等在这方面作了大量工作，证实此法不仅可用于细菌分类学的研究，而且也可应用于细菌检验等实际工作。现在，用高效毛细管柱已能从细菌细胞中分离出50多种不同的脂肪酸成分，一株典型的革兰氏阴性菌可含12—20种脂肪酸。许多菌株特别是在分类学上关系密切的类属菌株之间的差别，往往不在有无某一(或某些)特征成分的定性差异上，而是反映在各脂肪酸成分的定量差异上。早期用目测法比较各被试菌株的细胞脂肪酸 GC 图，解释它们之间的相互关系，用言语描述或列表表示，困难颇多，不易做到客观、定量，更无法使本方法计算机化乃至自动化。

系统聚类分析是统计模式识别的重要方法之一，现已广泛应用于化学模式识别。我们可以把细菌的细胞脂肪酸 GC 图看作是该细菌在特定条件下的一种化学模式，图中关于色谱峰的定量数据即是该模式的若干可以测量的特征。因此，根据细胞脂肪酸 GC 分析而进行的细菌分类鉴定问题，也就变为对细菌细胞脂肪酸模式进行识别的化学模式识别问题。Jantzen 等[2]曾用改进的 Yule 相关系数和系统聚类分析中的不加权的对-组平均法（即类平均法），对奈瑟氏菌（*Neisseria*）和莫拉氏菌（*Moraxella*）的细胞脂肪酸 GC 数据进行分析，取得较好效果。

我们在用全细胞脂肪酸气相色谱法鉴别莫拉氏菌和嗜肺军团杆菌（*Legionella pneumophila*）的实验研究中，用两种相似系数、8种系统聚类分析算法识别被试菌株的细胞脂肪酸模式，得到了初步结果。

本文于 1987 年 2 月 4 日收到。
* 中国科学院科学基金资助的课题。

材料与方法

（一）实验菌株

莫拉氏菌及其类属菌 34 株。其中国际标准株和参考株 12 株：*M. osloensis* 19976 (ATCC); *M. phenylpyruvica* 23333 (ATCC), 30004 (CMCC); *M. nonliquefaciens* 19975 (ATCC); *M. liquefaciens* 17952 (ATCC); *M. bovis* 30002 (CMCC); *M. lacunata* 30001 (CMCC); *M. branhamella* cata. 25238 (ATCC); *Acinetabacter calcoaceticus* 30101 (CMCC); *Kingella kingae* 23330 (ATCC); 我国新种代表株 *M. amylolytica* 8057; *M. nanchang* 8052。待定株 225、157、325、188、43、8062、121、8109、8103、HZ、10B、E21、Qiu、176、41、Liu、8104、64004、174、58B、8059、五-71。以上菌株由中国医学细菌保藏管理中心莫拉氏菌专业实验室提供。

嗜肺军团杆菌 13 株：L. *pneumophila* 1-8 型(美国 CDC)、NJ 8331 (南京军区总医院)，由军事医学科学院微生物流行病研究所提供；*L. pneumophila* (1)、(6)、LDB 4a 和 4b 由总后勤部卫生部直属防疫队提供。

（二）细菌细胞脂肪酸 GC 图的获得

细菌细胞脂肪酸的气相色谱分析使用 Pekin-Elmer SIGMA 115 型气相色谱仪，附加 3390A 型电子积分仪 (Hewlett-Packard 公司)，同步打印出百分峰高和百分峰面积。

关于菌株培养、细菌全细胞脂肪酸甲酯的制备以及气相色谱分析工作条件等，详见参考文献[3]。

（三）系统聚类分析

1.建立原始数据矩阵：选取被试菌株细胞脂肪酸 GC 图的 $C_{10:0}$ 至 $C_{24:0}$ 段进行分析，计算其中各成分峰的相对保留时间，参照样品和标准品的共色谱图 (co-chromatogram)，确定每个峰的对应位置。并根据积分仪给出的峰高百分数，制成各被试菌株的细胞脂肪酸成分表（见参考文献[3]），此即为供聚类分析用的原始数据矩阵。

2.计算各菌株间的相似系数：设菌株 i 和 j 的峰 k 的峰高百分值分别为 x_{ik} 和 x_{jk}，则菌株 i 与 j 之间的欧氏距离系数 D_{ij} 为：

$$D_{ij} = \left[\frac{1}{m}\sum_{k=1}^{m}(x_{ik} - x_{jk})^2\right]^{1/2},$$

它们的指数相关系数为：

$$R_{ij} = \frac{1}{m}\sum_{k=1}^{m}\exp\left[-\frac{3}{4}\cdot\frac{(x_{ij} - x_{jk})^2}{S_k^2}\right]。$$

式中，m 为被试菌株的色谱峰数，S_k 为色谱峰 k 的标准差。

3.系统聚类算法：根据类与类之间的距离的定义不同，系统聚类分析一般可有 8 种算法，即最短距离法、最长距离法、平均距离法、可变法、中间距离法、类平均法、重心法和离差平方和法。这 8 种算法的计算类间距离的递推公式可统一成如下形式(其中 $G_r = G_p \cup G_q$)：

$$D_{kr} = \alpha_p D_{kp} + \alpha_q D_{kq} + \beta D_{pq} + \gamma|D_{kp} - D_{kq}|。$$

式中，D_{kr} 为类 G_p 和 G_q 合并成 G_r 后某一类 G_k ($K \neq p, q$) 与 G_r 间的距离，D_{kp} 为类 G_k 与 G_p 间的距离，D_{kq} 为类 G_k 与 G_q 间的距离，D_{pq} 为类 G_p 与 G_q 间的距离。参数 α_p、α_q、β 和 γ 依算法不同而取不同的值[4]。用中间距离法和重心法时，相似系数为欧氏距离系数平方，用离差平方和法时，为欧氏距离系数平方之半。

计算机程序用 BASIC 语言编写，在 VAX 11/780 小型机上运行，稍作修改也可以在微型机（如 H-89 机等）上运行。

结　果

图 1—8 是用欧氏距离系数、8 种系统聚类分析算法对 47 株被试菌株的全细胞脂肪酸 GC 图进行聚类分析而得到的树状谱。最长距离法把这些菌株分为 6 群（图 2）。12 株嗜肺军团杆菌聚为一群，虽然它们是由两个实验室用不同的培养基培养的，但它们仍能聚于一群，这与常规鉴定的结果完全相符。我国莫拉氏菌专业实验室分离、收集的 9 株莫拉氏菌新种 *M. amylolytica* 聚为一群；该实验室分离收集的、另 2 株莫拉氏菌新种 *M. nanchang* 与 *M. lacunata* 30001 和 *K. kingae* 23330 聚在一起。这两个新种与其他莫拉氏菌参考株有许多共同特征，但又存在一定差别。这一

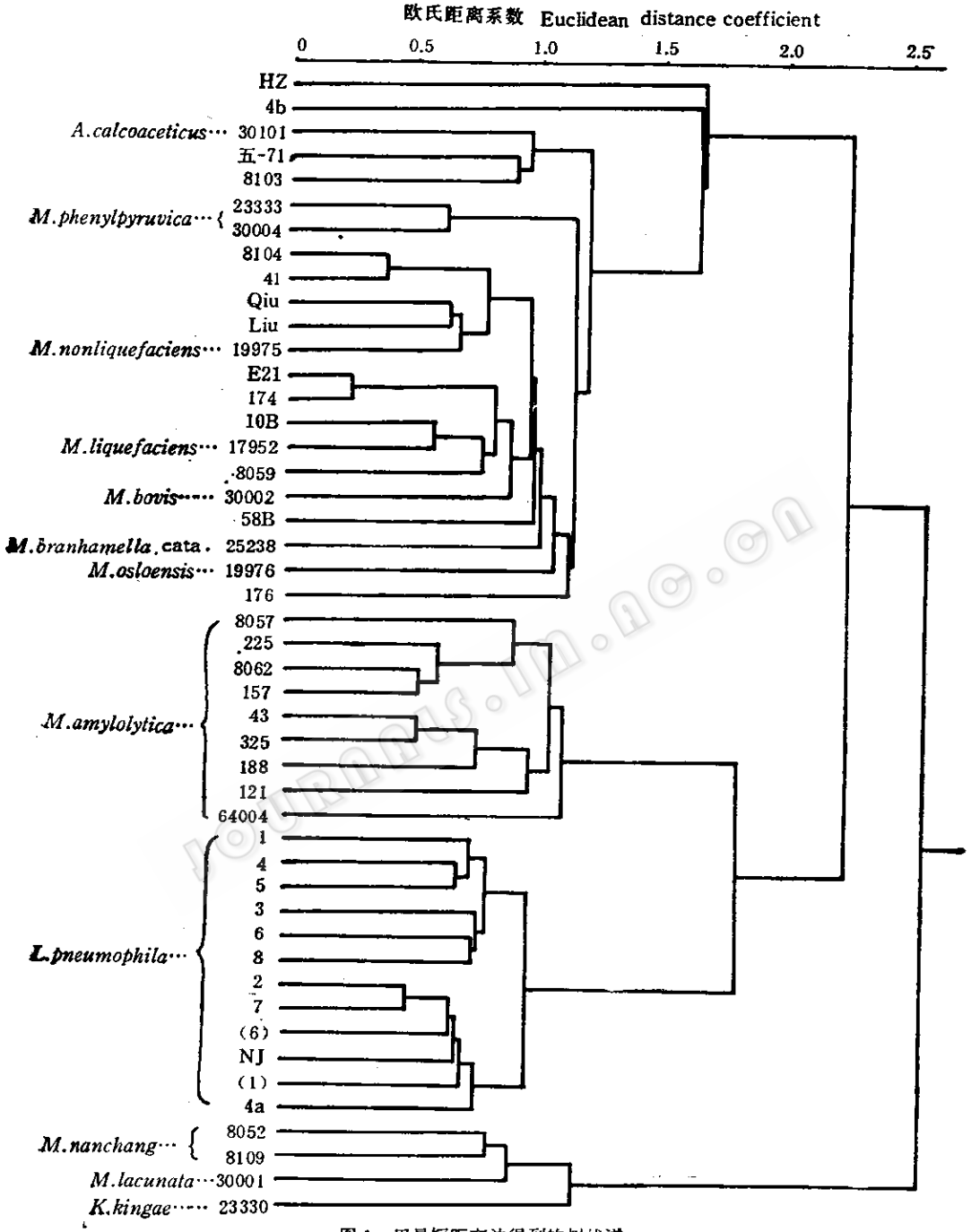

图 1 用最短距离法得到的树状谱

Fig. 1 The dendrogram obtained by the minimum method

点在本树状谱上得到充分反映，为确定它们的分类学位置提供了新的参考依据。其余 8 株莫拉氏菌参考株分属于两个群，多数待定株的聚类结果与常规鉴定结果相符，但有一些初步定名为 *M. osloensis* 的待定株 176、Qiu、8104、E21、41、Liu 和

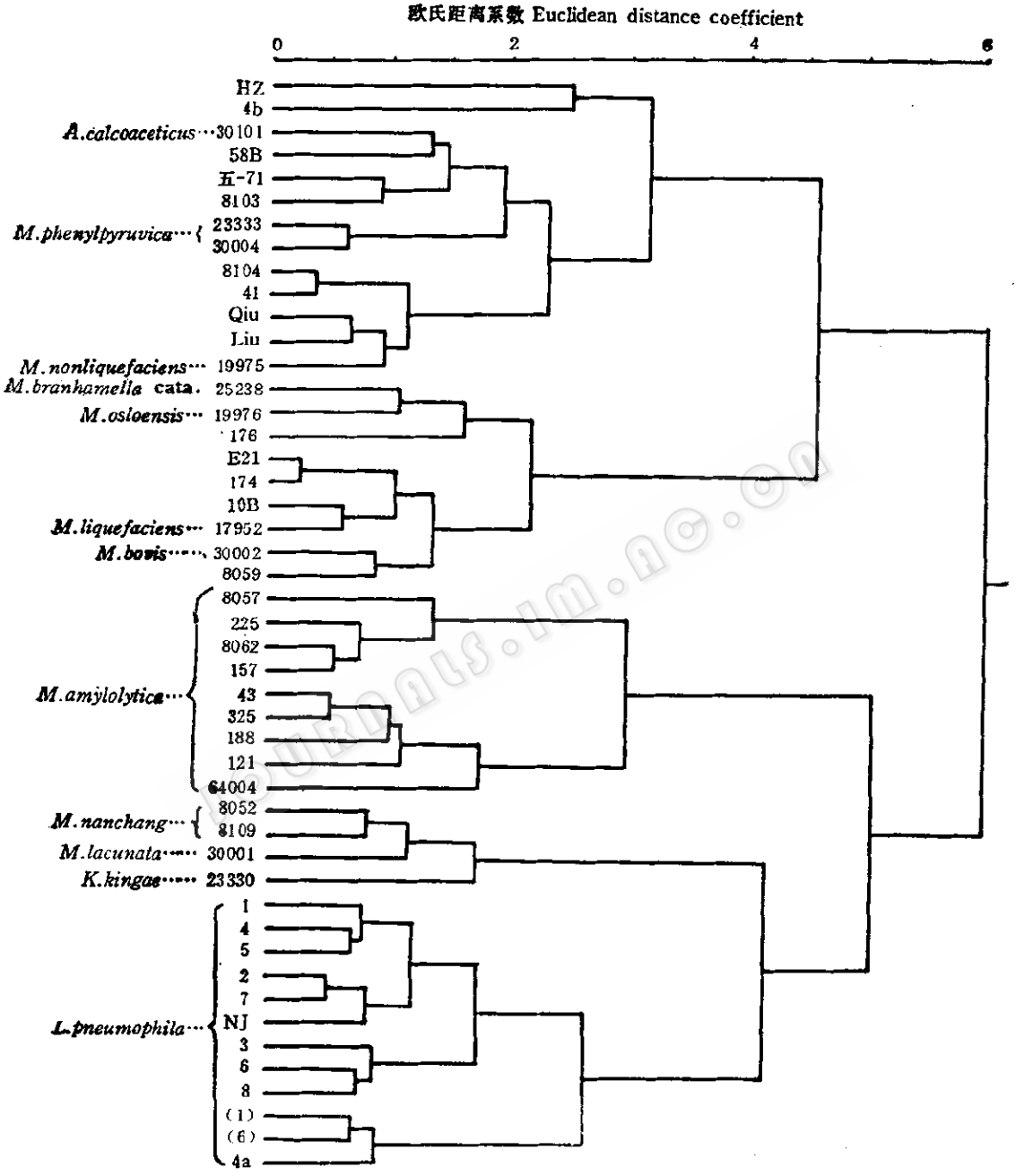

图 2　用最长距离法得到的树状谱

Fig. 2　The dendrogram obtained by the maximum method

10B 被分散在这两个群中，它们的常规试验结果也存在一定差异。从树状谱中还可以看出，初步定名为假单胞菌的 HZ 株确不属于莫拉氏菌，分离嗜肺军团杆菌时得到的 LDB 4b 不是嗜肺军团杆菌。它们的分类学位置有待进一步研究确定。

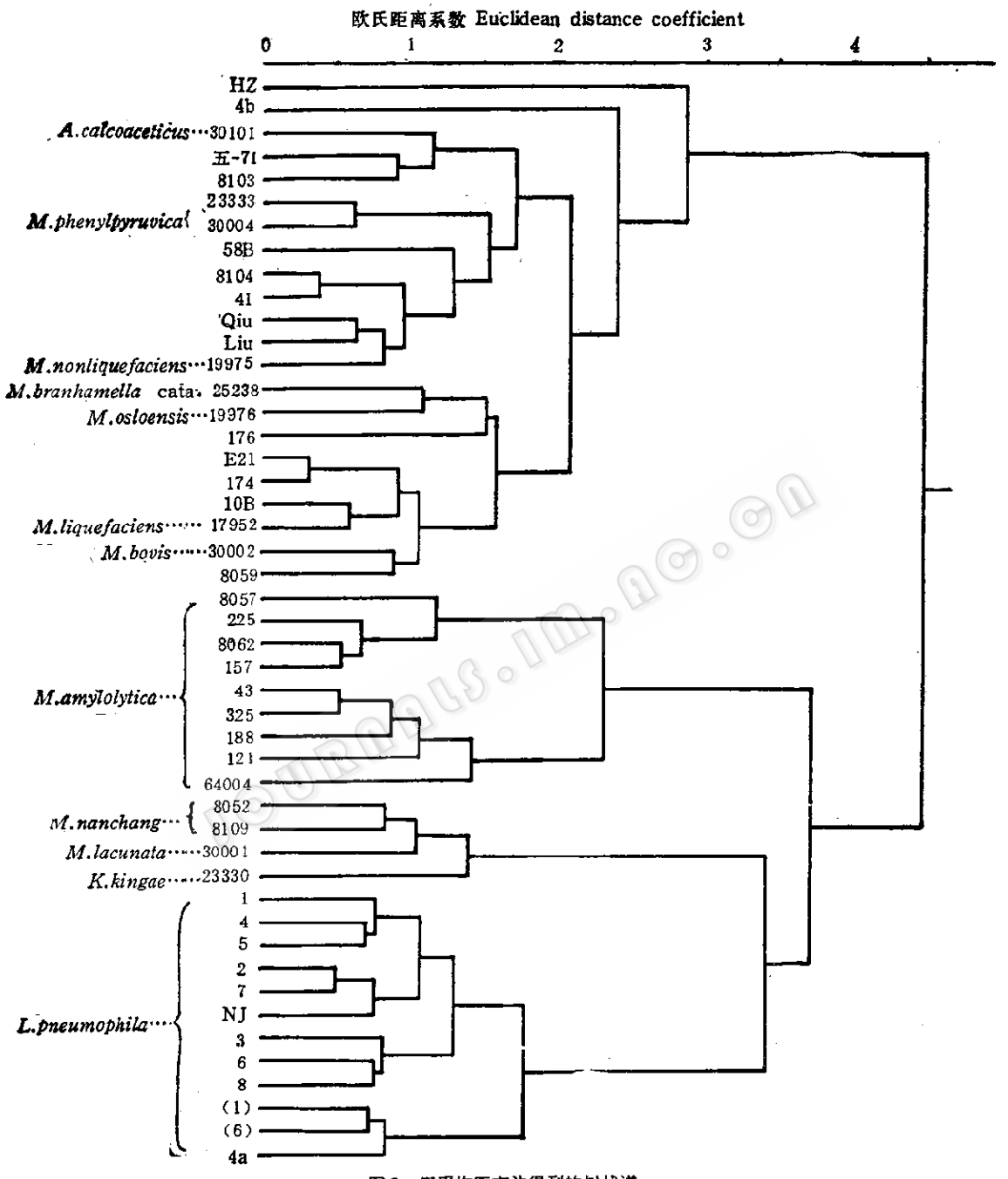

图3 用平均距离法得到的树状谱

Fig. 3 The dendrogram obtained by the average method

平均距离法、可变法（$\beta = -0.25$）和类平均法的聚类结果（图3、4和5）与最长距离法的结果基本一致，各对应群内所包含的菌株均相同，仅个别菌株在群内的位置略有区别，如菌株58B。最短距离法对菌株间的微小差异较不敏感，两个莫拉氏

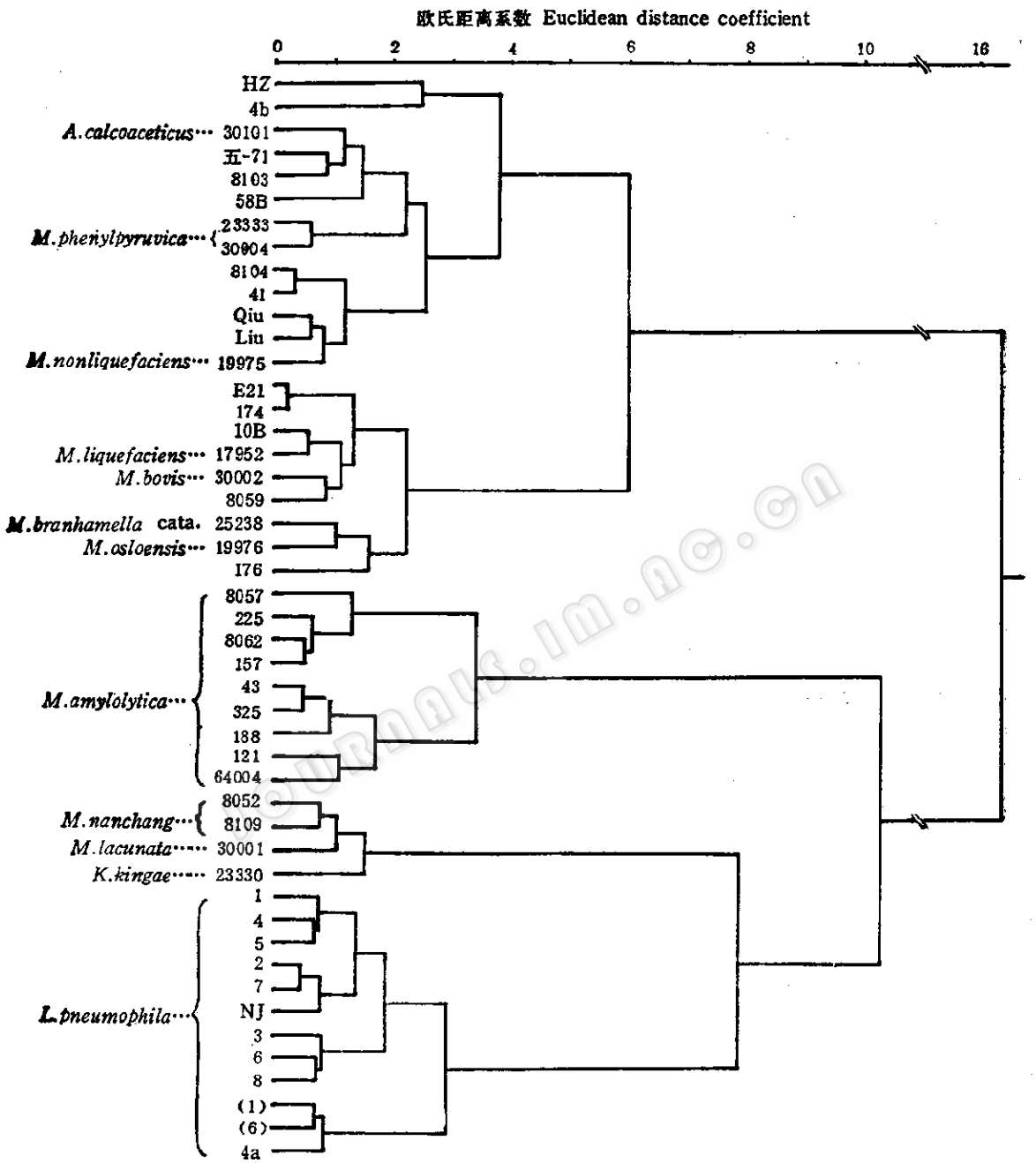

图 4　用可变法（$\beta = -0.25$）得到的树状谱

Fig. 4　The dendrogram obtained by the flexible method

菌新种和嗜肺军团杆菌这三群分得较好（图 1），其余菌株的分类效果似不如上述四种算法。中间距离法和重心法的聚类结果（图 6 和 7）相同，它们的一个共同缺点是树状谱中可出现逆转，即上一级聚合高度反而小于下一级聚合高度的现象。本例

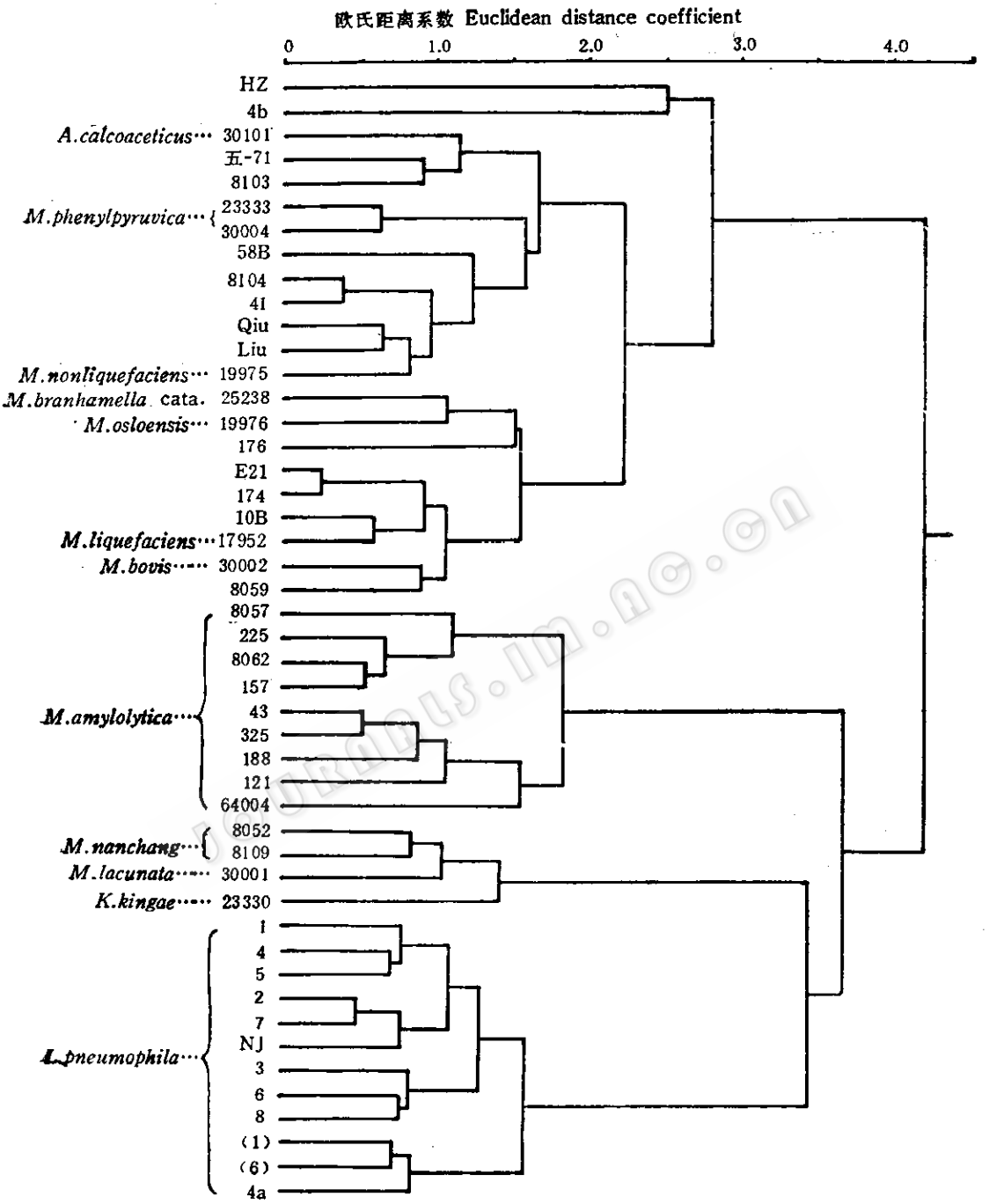

图 5 用类平均法得到的树状谱

Fig. 5 The dendrogram obtained by the group average method

中出现 5 处逆转，这种非单调性打乱了系统聚类过程中的等级关系，给合理解释聚类结果带来困难。离差平方和法的类间距离的跨度太大，本例中竟达 4 个数量级(图 8)，给树状谱的绘制带来不便，在这里不得已改用对数坐标表示。

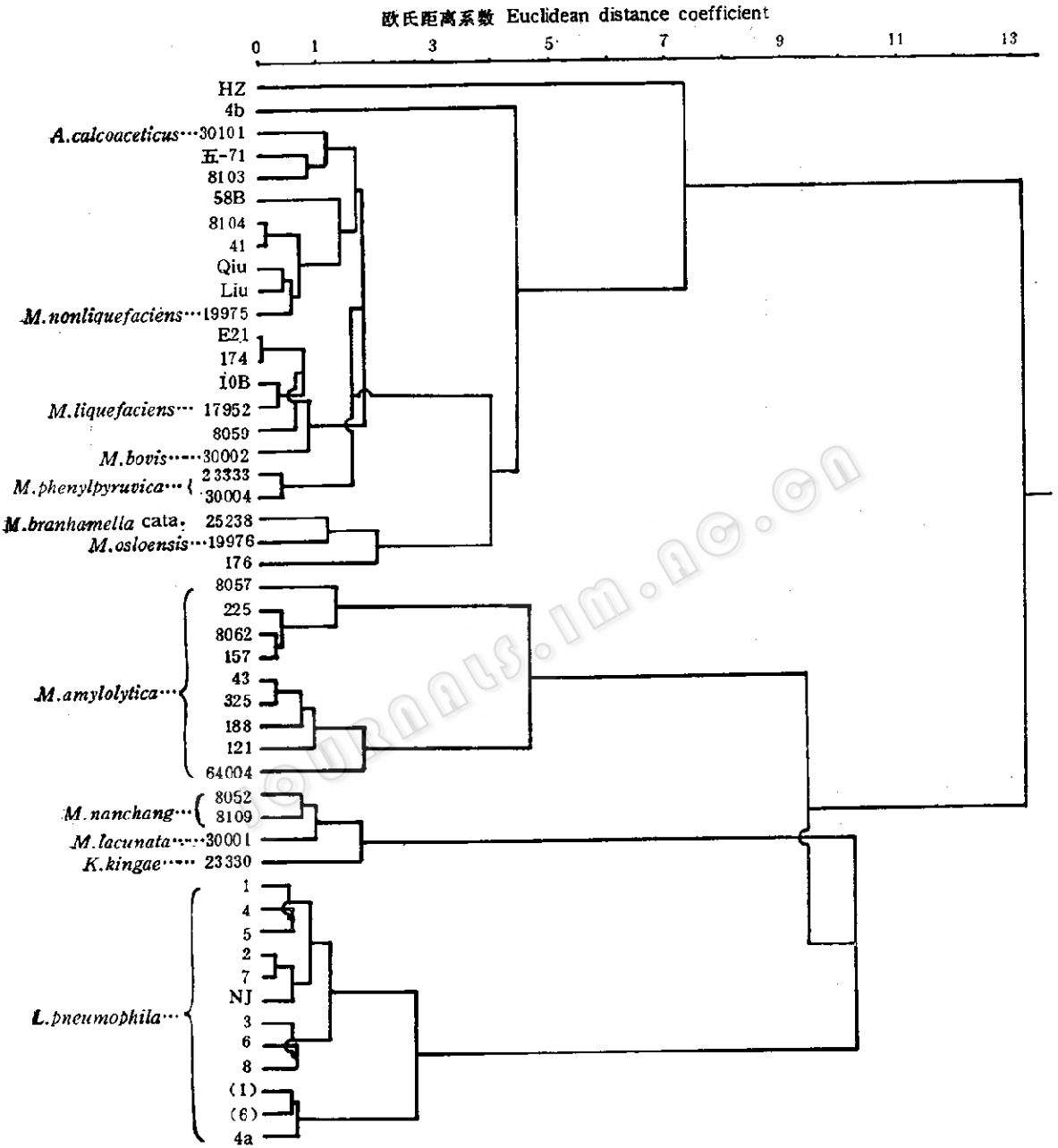

图 6　用中间距离法（$\beta = -0.25$）得到的树状谱

Fig. 6　The dendrogram obtained by the median method

采用不同的相似系数对菌株聚类结果的影响甚大。用指数相关系数为相似系数，对 47 株被试菌株进行聚类分析所得树状谱与用欧氏距离系数所得者明显不同，与常规鉴定结果比较，符合程度也较差。其原因有待进一步分析。

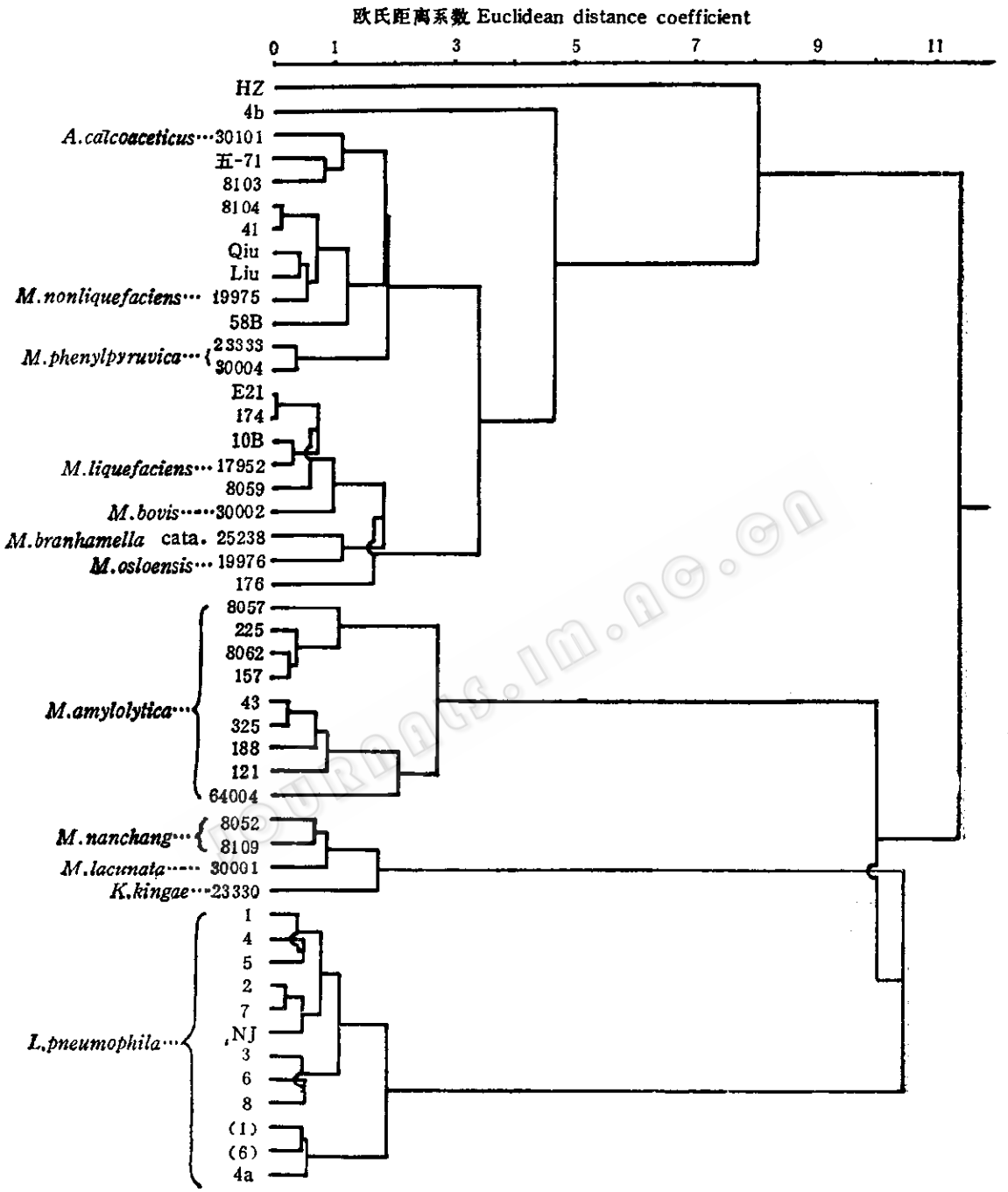

图 7 用重心法得到的树状谱

Fig. 7 The dendrogram obtained by the centroid method

讨 论

1. 建立准确可靠的原始数据矩阵极为重要，这是聚类分析的基础之所在。用气相色谱法分析细菌细胞脂肪酸成分时，由于种种原因，各菌株样品对应成分峰的保

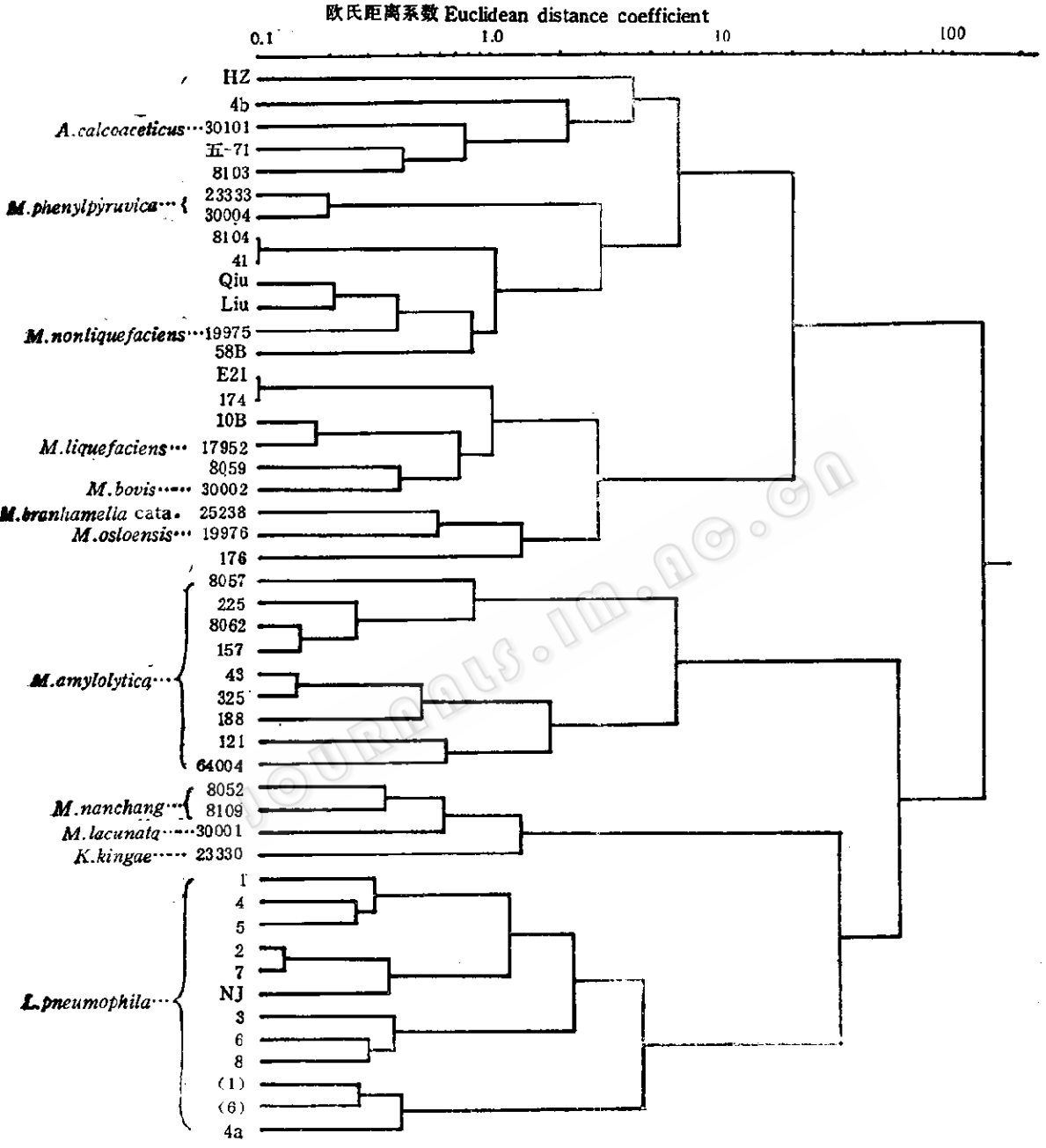

图 8　用离差平方和法得到的树状谱

Fig. 8　The dendrogram obtained by the ward's method

留时间很难完全一致，实验日期相隔越久，这种不一致性可能就越严重。在对这样的色谱图作聚类分析时，首先必须准确地定出各菌株样品的每个成分峰的对应位置，制定出能正确反映真实情况的原始数据表。在实验过程中，我们发现仅选一个"参

考峰”，计算其余各峰相对于此峰的相对保留时间是不能满足要求的，特别是那些离参考峰较远的成分峰，误差更大。为此，我们在分析细菌样品的同时，还将脂肪酸标准品加入细胞脂肪酸样品中，作共色谱。以所得共色谱图为参照，反复比较、判别，把各被试菌株的每个成分峰的对应位置确定下来，取得了较好的效果。

2. 在本文的初步结果中，用欧氏距离系数作相似系数可较好地表示被试细菌细胞脂肪酸模式的相互关系。不过，在细菌细胞脂肪酸 GC 图中，有的峰很强也很稳定，但它们对鉴别被试菌株未必能提供较多的信息；而有的峰虽小，但具有很强的特征性，在分类学上有重要意义。欧氏距离系数不能反映谱峰的这些特点，影响其聚类效果。若对欧氏距离系数作适当的加权处理[5]，则有可能进一步改善其聚类效果。

最短距离法具有较强的空间压缩性，使被试菌株有链接聚合的趋势；中间距离法和重心法有严重的非单调性，使聚类树状谱中出现逆转现象，影响对结果的解释。由于这些原因，这三种方法现已很少使用。一般认为，类平均法既是空间保持的，又具单调性，是比较理想的聚类分析方法，我们的初步结果也证实了这一点。本文的结果还表明，最长距离法用于细菌细胞脂肪酸模式识别，也取得了较好的效果，本法计算类间距离的递推公式中的四个参数是设定的，在聚类过程中不再改变，故较易于计算机编程。在可变法中，参数 β 称为聚集强度系数，其聚类结果随 β 的取值不同而不同。β 取正值时，聚类是空间压缩的；$\beta=0$ 时，是空间保持的；β 取负值时，是空间扩张的，且 β 的绝对值越大，扩张性能也越强。但扩张不宜过强，否则会出现一些不合理的分类。一般取 $\beta=-0.25$，本文中即取此值。

3. 从本文的初步结果看，采用合适的相似系数和系统聚类分析算法，对细菌细胞脂肪酸模式进行聚类分析，至少能把被试菌株鉴别至属的水平。至于属以下的水平，看来情况比较复杂。如在莫拉氏菌属内，虽能把我国分离的 2 个新种与目前该属的主要标准株分开，前两者之间也能区分，且与常规结果符合较好，但在后者中，有些标准株与类属菌之间不能明确区分，一些待定株的聚类结果与常规鉴定结果不符。这一方面是因为莫拉氏菌的变异性较强，存在许多不典型菌株和过渡型，另一方面是因为本实验研究的化学分析程序主要集中在分析细胞中的直链脂肪酸，其他具有重要分类学意义的脂肪酸如羟基脂肪酸尚未涉及。本文的结果仅是初步的，不论在相似系数还是聚类分析方面的工作都有待深入。

在军团杆菌中，可进一步把两个实验室在不同培养基中培养的菌株细分开。不论其血清型如何，同一实验室培养的菌株先归为一类，这说明培养基及其他培养条件有可能影响细菌的细胞脂肪酸组成，进而影响最终的聚类结果。因此，在采用细胞脂肪酸气相色谱法鉴别细菌时，严格控制从菌株培养到细胞脂肪酸分析的全部实验条件一致，是十分重要的。

参 考 文 献

[1] Abel, K. et al.: *J. Bacteriol.*, 85: 1039—1044, 1963.

[2] Jantzen, E. et al.: *Acta Path. Microbiol. Scand. Sect. B*, 82: 767—779, 1974.

[3] 周方泰等：中国科学，B 辑，第 10 期：1051—1058，1986。

[4] 方开泰、潘恩沛：《聚类分析》，地质出版社，北京，1982。

[5] Wieten, G. et al.: in Gas Chromatography/Mass Spectrometry Applications in Microbiology (ed. by G. Odham et al.), Plenum Press, New York, pp. 335—379, 1984.

APPLICATION OF THE HIERARCHICAL CLUSTERING ANALYSIS TO RECOGNIZATION OF BACTERIAL WHOLE-CELL FATTY ACID PATTERNS

Zhu Houchu

(Institute of Biotechnology, Academy of Military Medical Science, PLA, Beijing)

Zhou Fang　Tang Guangjiang

(Institute of Microbiology and Epidemology, Academy of Military Medical Science, PLA, Beijing)

Hierarchical clustering analyses have been done on whole-cell fatty acid patterns obtained from 34 strains of *Moraxella* and its related bacteria and 13 strains of *Legionella pneumophila* by capillary gas chromatography, using combinations of the Euclidean distance coefficient and the exponential correlation coefficient with eight strategies of hierarchical clustering. A comparison between the dendrograms obtained by these strategies has been made. The results showed that there were defined discriminations between *Moraxella* and *L. pneumophila*, and between two new species isolated in China and the currently established referance strains in *Moraxella*. The Euclidean distance coefficient was more effective than the exponential correlation coefficient. The maximum method and the group average method had the advantage of other strategies.

Key words

Hierarchical clustering analysis; Whole-cell fatty acid patterns; *Moraxella; Legionella pneumophila*