185—188 21239(15)

第12卷第2期 中 国 病 毒 学 Vol.12 No.2
1997年6月 VIROLOGICA SINICA Jun. 1997

不同系统聚类算法对昆虫包涵体病毒模式的识别

刘进学 朱湘民 汤显春
张世敏 陈 华* 夏祥明

(中国科学院武汉病毒研究所,武汉 430071)

Q965.8

关键词 病毒模式, 系统聚类分析, 分类鉴定 昆虫病毒,包涵体病毒

随着分析微生物学技术的不断发展,毛细管裂解色谱法(Pyrolysis Gas chromatography,简称PyGC)业已用于鉴定微生物[1-4]和分析鉴别化学成分、结构极为复杂的生物大分子[5],并证明了利用PyGC分析鉴定昆虫包涵体病毒的可行性[6]。系统聚类分析是统计模式识别的重要方法之一,已广泛应用于化学模式的识别。我们使用欧氏距离系数的8种常规系统聚类算法,对48株典型昆虫包涵体病毒株,即26株核型多角体病毒(NPV)、13株颗粒体病毒(GV)和9株质型多角体病毒(CPV),用PyGC制备的气象色谱(GC)图进行聚类分析,并对不同聚类算法得到的树状谱进行比较,取得了较好的结果。

材料和方法

1 供试毒株

NPV26株:(1)EpNPV-SH,(2)EpNPV-to,(3)EpNPV-F,(4)EpNPV-An,(5)EpNPV-Xi,(6)EpNPV-Si,(7)EpNPV-En,(8)EaNPV-Ji,(9)EaNPV-Xi,(10)EaNPV-He,(11)EaNPV-S,(12)EpNPV-Me,(13)EfNPV-Gu,(14)BsNPV-Xi,(15)BsNPV-Hu,(16)BtNPV-Sa,(17)BtNPV-Me,(18)EcNPV-G,(19)HaNPV-W,(20)PiNPV-W,(21)PiNPV-Ti,(22)TsNPV-Ca,(23)HcNPV-L,(24)PcNPV-Gu,(25)PiNPV-Z,(26)HaNPV-Yu;GV13株,(27)PrWd,(28)PrGV-G,(29)AsGV-Bei,(30)PrGV-C,(31)ApGV-An,(32)MsGV-Hu,(33)PrGV-Zh,(34)PiGV-Sn,(35)CfGV-Cn,(36)CfGV-Si,(37)CfGV-Hu,(38)PrGV-W,(39)PrGV-S;CPV9株,(40)BmCPV-S,(41)BmCPV-Sn,(42)BmCPV-T,(43)BmCPV-Sh,(44)BmCPV-Zh,(45)BmCPV-An,(46)BmCPV-Au,(47)BmCPV-Ch,(48)BmCPV-Ak。

以上毒株均由中国菌种保藏委员会普通病毒保藏中心提供。

2 毒株GC图的获得

包涵体病毒的增殖、纯化、裂解气相色谱分析条件,参见文献[6,7]。48株病毒GC图的例图见图1。

3 系统聚类分析

3.1 建立原始数据矩阵 从每个毒株样品的3张重复GC图中选出一张,分别在保留时间(单位:min)为5.075、5.525、6.658、6.800、7.092、7.658、8.333、10.758、12.685、13.250、14.242、15.125、15.625、16.783、16.975、18.167、18.467、20.642和23.850处,确定每株在色谱数据处理机上给出的峰高百分数,制成的成分表,

收稿日期:1995-11-22,修回日期:1996-12-02

* 湖北省公安厅刑侦科学研究所

即为供各被试毒株聚类分析用的原始数据矩阵。

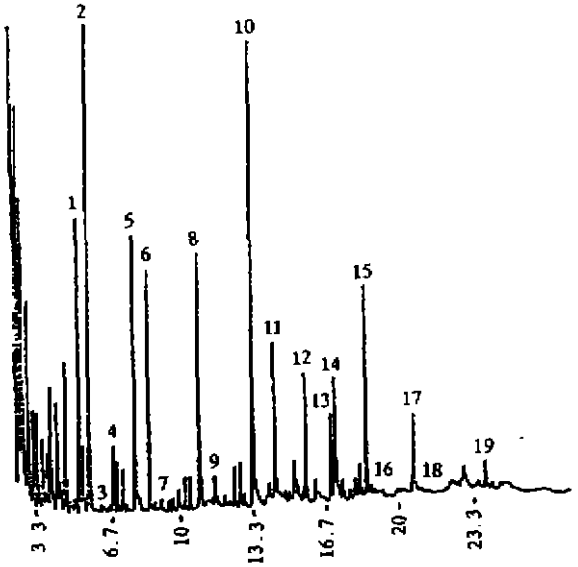

图1 昆虫包涵体病毒GC图

Fig 1 The gas chromatogram of insect inclusion body viruses

3.2 计算各毒株间的相似系数 设毒株i和j的峰K的峰高百分值分别为Xik和Xjk，则毒株i与j间的欧氏距离系数Dij为：$Dij=[\frac{1}{m}\sum_{k=1}^{m}(X_{ik}-X_{jk})^2]^{1/2}$，它们的指数相关系数为：$Rij=\frac{1}{m}\sum_{k=1}^{m}\exp[-\frac{3}{4}\times\frac{(x_{ij}-x_{jk})^2}{sk^2}]$。式中，m为被试毒株的色谱峰数，SK为色谱峰K的标准差。

3.3 系统聚类算法 采取欧氏距离系数的8种算法，即最短距离法、最长距离法、类平均法、重心法、中间距离法、离差平方和法、平均距离法和可变法。这8种算法计算类间距离的递推公式可统一成如下形式(其中Gr=GpUGq)：$D_{kr}=\alpha_p D_{kp}+\alpha_q D_{kq}+\beta_{pq}+\gamma|D_{kp}-D_{kq}|$。

式中，Dkr为类Gp和Gq合并成Gr后某一类Gk(K≠p，q)与Gr间的距离，Dkr为类Gky与Gp间的距离，Dkr为类Gk与Gq间的距离，Dpq为类Gp与Gq间的距离。参数αp、αq、β和γ依算法不同而取不同的值[8]。

3.4 计算机处理 计算机程序用汇编语言完成，在COPAQ386/20e微机及其兼容机上运行。

结果与讨论

本试验从包涵体病毒的增殖、样品的纯化到色谱分析等方面都力求规范一致，并尽量减少前后试验的时间间距。以便有效地提高毒株样品各对应成分峰位置的吻合度，增加"原始数据矩阵"的真实和可靠性是保证聚类分析的关键。

在8种算法中，以平均距离法、可变法(β=0.00)和类平均法的树状谱较理想。它们同属空间保持的，又具有单调性。在欧氏距离0.7以下，48株包涵体病毒明确聚合为四群。即7株NPV、6株GV和6株NPV聚合成第一群；由7株GV聚合为第二群；13株NPV聚合为第三群；9株CPV聚合为第四群。再由第一、二、三群依次聚合成杆状病毒属，最后与第四群胞质型多角体病毒(属)聚合，得到48个毒株的完整树状谱(图2)。这些树状谱明确展示出昆虫包涵体病毒间的亲疏关系。有些毒株在各树状谱中均保持着彼此紧靠和近距离聚合的稳定性，这些毒株有可能是近似的。其中(38)PrGV-W和(39)PrGV-S经血清学、多肽分析和DNA酶切图谱鉴定都证明二者系同一种病毒[7]。

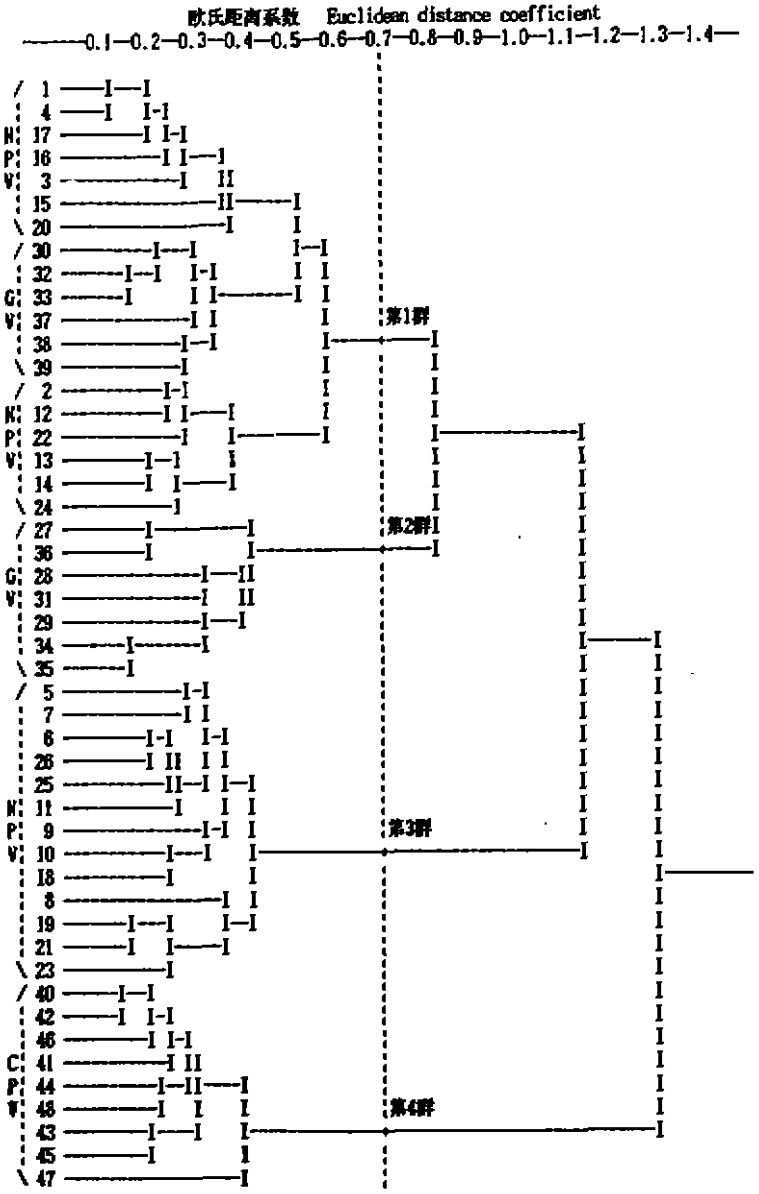

图 2　用平均距离法和可变法(β=0.00)得到的树状谱

Fig. 2　The dendrogram obtained by the average method and the flexible method

用欧氏距离系数进行聚类虽然得到较好结果,然而也存在某些不足。在毒株 GC 图上有的峰很强且稳,但为鉴别毒株未必能提供较好信息;有的峰虽小,却有很强的特征性,在分类上有重要意义(图 1)。如对其作适当的加权处理,有可能进一步改善聚类效果。在欧氏距离的 8 种算法中,某些算法存在着严重缺陷。如最短距离法、离差平方和法和可变法(β=0.25)的树状谱均出现空间压缩,甚至出现毒株连接聚合的趋势,它们对毒株间的微小差别不够敏感,给分组分群造成困难。中间距离法和重心法在聚类过程中出现非单调性,产生上一级聚合度反

而低于下一级聚合度的逆转现象，打乱了系统聚类的等级关系，影响对结果的解释。这种现象和有关报道情况一致[8]。由于上述原因，这些方法目前人们已很少使用。最长距离法和可变法($\beta=0.25$)在聚类过程中，违背了先属内后属间的一般规律，导致昆虫包涵体病毒分类的属间关系混乱。

正确的聚类算法，可反映毒株与毒株间的亲疏关系，它为再认识某些现有病毒和鉴定新毒株提供了佐证。昆虫包涵体病毒的系统聚类分析，实质上是这些毒株的化学模式识别，这与过去以生物学为主的传统分类鉴定有其较大距离。随着这方面工作的不断深入，可能系统聚类分析对昆虫病毒分类鉴定会起到积极作用。

参 考 文 献

1 Oyama V.I. Use of gas chromatography for the detection of life on mars. Nature, 1963, 200: 1058

2 Reiner E. Pyrolysis gas liquid chromatography studies for the classification of mycobacterial. Am Rev Respir, Dis. 1969, 99: 750－759

3 Reiner E. Rapid characterization of salmnella organisms by menas of pyrolysis－gas－liquid chromatography. Anal. Chem. 1972, 44: 1058～1061

4 周方，王菊英，王给山等．裂解气相色谱法区分杆菌及其类属菌 科学通报，1984，29(22)：1934～1397

5 朱相民，赵姬勇，罗清修等．降解性质粒 DNA 的裂解气液色谱分析．微生物学报，1988，28(1)：62

6 朱相民，汤曼春．用裂解气液色谱鉴定昆虫包涵体病毒的初步研究．病毒学报，1987，3(4)：355～360

7 朱相民，汤曼春，刘进学等．PyGC 对昆虫包涵体病毒的模式识别．分析测试学报，1995，14(6)：37～42

8 朱厚础，周方、唐光江．系统聚类分析在细菌全细胞脂肪酸模式识别中的应用．微生物学报，1987，27(4)：306～317

The Eight Strategies of the Hierarchical Clustering Analysis for Recognition of Patterns of the Insect Inclusion Body Viruses

Liu Jinxue Zhu Xiangmin Tang Xianchun *et el*

(*Wuhan Institute of Virology, Academia Sinica, Wuhan* 430071)

Hierarchical clustering analysis have been done on 48 strains of the representative insects inclusion body viruses (26 strains of nuclear polyhedrosis viruses, 13 strains of granulosis viruses and 9 strains of cytoplasmic polyhedrosis viruses) by capillary gas chromatography, using the eight strategies of hierarchical clustering of Euclidean distance coefficient. A comparison between the dendrograms obtained by these strategies has been made. The results showed that there are definite discriminations between baculovirus group (genus) and cytoplasmic polyhedrosis virus group (genus). The relationships among the genera, groups, subgroups and strains could clearly be unfolded. The similarities and differences among strains also be distinguished. The average method, flexible method ($\beta \doteq 0.00$), and group average method in eight strategies of hierarchical clustering had the advantage of other strategies.

Key words Pattlern of virus, Hierarchical clustering analysis, Classification