技术与方法

# 模糊聚类分析法在昆虫杆状病毒分类中的应用

李 虹
(中国科学院武汉数学物理所)

刘明富
(中国科学院武汉病毒研究所)

**摘要** 采用一种新的数学方法——模糊聚类分析法,对已知包涵体蛋白 N-末端序列的 10 种昆虫杆状病毒进行分类,其结果与 Fitch 方法的结果完全一致。这种分类方法并不只是对各序列进行简单的类别划分,而且也表示了各类间相似的程度及在某一类内各序列间的进化关系。

**关键词** 模糊聚类分析法;杆状病毒;分类

昆虫杆状病毒科 A 亚群-核多角体病毒和 B 亚群-颗粒体病毒被广泛用作杀虫剂,近年来,又作为外源基因的高效表达载体,越来越受到重视[1]。在进化上,具有相同功能的多角体蛋白或颗粒体蛋白(统称包涵体蛋白)由单拷贝基因编码,表现出很强的保守性。根据这个性质,很多杆状病毒的包涵体蛋白基因已被定位,并分析了其核苷酸序列,尤以苜蓿银纹夜蛾核多角体病毒研究得最为清楚[1-3]。以这种病毒包涵体蛋白序列为标准,Rohrmann[2] 按 Fitch 方法建立了 10 种包涵体病毒(8 种 NPV, 2 种 GV)的系统进化树,为杆状病毒起源与进化的研究提供了有用的资料。本文采用另一种新的数学方法——模糊聚类分析法对上述 10 种包涵体病毒进行分类,结果报道如下。

## (一) 包涵体病毒及包涵体蛋白 N-末端残基序列

从已有资料表明,包涵体蛋白 N-末端变异性最大[2],本文即从下述 10 种病毒包涵体蛋白 N-末端入手,采用模糊聚类分析方法对其进行分类。这 10 种包涵体病毒分别是:

1. 苜蓿银纹夜蛾多粒包埋核多角体病毒 (*Autographa californica* multi-nucleocapsid nuclear Polyhedrosis virus, AcMNPV)

2. 大蜡螟多粒包埋核多角体病毒 (*Galleria mellonella* MNPV, GmMNPV)

3. 家蚕单粒包埋核多角体病毒 (*Bombix mori* single-nucleocapsid nuclear polyhedrosis virus, BmSNPV)

4. 舞毒蛾多粒包埋核多角体病毒 (*Lymantria dispar* MNPV, LdMNPV)

5. 黄杉毒蛾多粒包埋核多角体病毒 (*Orgyia pseudotsugata* MNPV, OpMNPV)

6. 沼泽大蚊单粒包埋核多角体病毒 (*Tipula paludosa* SNPV, TpSNPV)

7. 黄杉毒蛾单粒包埋核多角体病毒 (OpSNPV)

8. 大菜粉蝶颗粒体病毒 (*Pieris brassicae* granulosis virus, PbGV)

9. 粉纹夜蛾颗粒体病毒 (*Trichoplusia ni* GV, TnGV)

10. 松针黄叶蜂单粒包埋核多角体病毒 (*Neodiprion sertifer* SNPV, NsSNPV)

10 种病毒包涵体蛋白 N-末端 39 个氨基酸残基序列排列见图 1[2]。

## (二) 模糊聚类分析步骤[4,5]

1. 建立表示相似关系的矩阵 $R$:

设有 $m$ 个蛋白,$n$ 个氨基酸的序列矩阵为:

$$X=\begin{bmatrix} X_{11}, X_{12}, \cdots X_{1n} \\ X_{21}, X_{22}, \cdots X_{2n} \\ \cdots \\ X_{m1}, X_{m2}, \cdots, X_{mn} \end{bmatrix}$$

本文承蒙谢天恩教授、范文涛教授、吴远明副教授审阅并提出宝贵意见,特致谢。

| | 10 | 20 | 30 | 39 |
|---|---|---|---|---|
| 1. AcMNPV | M----PDYSY | PTIGRRTYVY | DNKYYKNLGA | VIKNAKRKK |
| 2. GmMNPV | .----.N... | ..... | ..... | ..... |
| 3. BmSNPV | .----.N... | N..... | .....G | L..... |
| 4. LdMNPV | .----KN..- | -.AL.K..... | .....T | ...Q...Q. |
| 5. OpMNPV | .----..... | ..... | .....S | ..... |
| 6. TpSNPV | .QYGVN..G. | E. NVDYPNAS | HAGL.DRSKQ | PYVDHD?YP |
| 7. OpSNPV | .---YTR... | N.SL..... | ..... | ..... |
| 8. PhGV | .-GYNRALR. | SKHE.T.C.I | ..QH..S... | .L.DV.H.. |
| 9. TnGV | .-GTNKSLR. | SRHN.T.C.I | ...HL.T..S | .LGDVRH.E |
| 10. NsSNPV | .---PNLAQG | YQ?SAKS.I. | .....G..D | I..S..??? |

图 1　10 种包涵体蛋白 N-末端序列("·"表示与 AcMNPV 相同,"-"表示缺失,"?"未确定)

对其中任意两个序列 $X_i$, $X_j$:

$$\begin{cases} X_i = X_{i1}, X_{i2}, \cdots, X_{in} \\ X_j = X_{j1}, X_{j2}, \cdots, X_{jn} \end{cases} \quad i, j = 1, 2, \cdots, m$$

定义

$$r_{ij} = \frac{\sum_{k=1}^{n} (X_{ik} \odot X_{jk})}{n} \qquad (1)$$

$r_{ij}$ 表示 $X_i$ 和 $X_j$ 的相似程度,其中:

$$X_{ik} \odot X_{jk} = \begin{cases} 1 & X_{ik} \text{ 与 } X_{jk} \text{ 相同} \\ 0 & X_{ik} \text{ 与 } X_{jk} \text{ 不同} \end{cases}$$

$$k = 1, 2, \cdots, n$$

显然 $0 \leqslant r_{ij} \leqslant 1$。本文中 $m = 10$, $n = 39$, 建立的模糊矩阵 $\underset{\sim}{R}$ 见表 1。

2. 将模糊矩阵 $\underset{\sim}{R}$ 改造为模糊等价关系 $\underset{\sim}{R}^*$:

上面所建矩阵 $\underset{\sim}{R}$ 只满足自反性和对称性,即为相似矩阵,不具备传递性。使用求 $\underset{\sim}{R}$ 的传递闭包的方法将 $\underset{\sim}{R}$ 改造为模糊等价关系 $\underset{\sim}{R}^*$。

$$\underset{\sim}{R} \to \underset{\sim}{R}^2 \to \underset{\sim}{R}^4 \to \cdots \to \underset{\sim}{R}^{2^l} = \underset{\sim}{R}^*$$

$$\underset{\sim}{R}^{2^i} = (r_{ij}^{(2^i)}) \quad i, j = 1, 2, \cdots, m$$

$\underset{\sim}{R}$ 的自乘公式为:

$$r_{ij}^{(2^i)} = \bigvee_{k=1}^{m} [r_{ki}^{(2^{i-1})} \wedge r_{kj}^{(2^{i-1})}] \quad i, j = 1, 2, \cdots, m$$

计算结果表明 $\underset{\sim}{R}^4 = \underset{\sim}{R}^8$, 从而 $\underset{\sim}{R}^4$ 为等价关系矩阵即 $\underset{\sim}{R}^* = \underset{\sim}{R}^4$ (见表 2)。

3. 作出动态聚类图对序列进行分类:

首先对任意 $\lambda \in [0, 1]$ 作 $\underset{\sim}{R}^*$ 的 $\lambda$-截矩阵 $\underset{\sim}{R}_\lambda^*$, 记为:

$$\underset{\sim}{R}_\lambda^* = ({}_\lambda r_{ij})$$

其中

$${}_\lambda r_{ij} = \begin{cases} 1 & r_{ij} \geqslant \lambda \\ 0 & r_{ij} < \lambda \end{cases}$$

得到如下动态聚类图(图 2)。在 $\lambda = 0.69$ 处截取,可将这 10 种序列分为四类:序列 1, 2, 5, 3, 7, 4, 为第一类, 10 为第二类, 8、9 为第三类, 6 为第四类。

4. 分类结果准确性验证——类内比较和类间比较:

任意选取序列 1, 10, 8, 6 分别为第一、二、三、四类的标准序列,按前面定义 (1) 式与所在类的所有序列进行比较,即求得每类内各

表 1　10 种包涵体病毒的包涵体蛋白相似关系的模糊矩阵

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | | | | | | | | | |
| 2 | 0.97 | 1.00 | | | | | | | | |
| 3 | 0.90 | 0.92 | 1.00 | | | | | | | |
| 4 | 0.74 | 0.77 | 0.77 | 1.00 | | | | | | |
| 5 | 0.97 | 0.95 | 0.90 | 0.74 | 1.00 | | | | | |
| 6 | 0.16 | 0.13 | 0.13 | 0.10 | 0.16 | 1.00 | | | | |
| 7 | 0.85 | 0.85 | 0.87 | 0.72 | 0.82 | 0.13 | 1.00 | | | |
| 8 | 0.46 | 0.46 | 0.41 | 0.38 | 0.44 | 0.08 | 0.46 | 1.00 | | |
| 9 | 0.36 | 0.36 | 0.33 | 0.31 | 0.38 | 0.05 | 0.36 | 0.69 | 1.00 | |
| 10 | 0.51 | 0.51 | 0.46 | 0.54 | 0.51 | 0.06 | 0.51 | 0.26 | 0.23 | 1.00 |

表 2 10 种包涵体病毒包涵体蛋白模糊等价关系矩阵 $R^*$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | | | | | | | | | |
| 2 | 0.97 | 1.00 | | | | | | | | |
| 3 | 0.92 | 0.92 | 1.00 | | | | | | | |
| 4 | 0.77 | 0.77 | 0.77 | 1.00 | | | | | | |
| 5 | 0.97 | 0.97 | 0.92 | 0.77 | 1.00 | | | | | |
| 6 | 0.16 | 0.16 | 0.16 | 0.16 | 0.16 | 1.00 | | | | |
| 7 | 0.87 | 0.87 | 0.87 | 0.77 | 0.87 | 0.16 | 1.00 | | | |
| 8 | 0.46 | 0.46 | 0.46 | 0.46 | 0.46 | 0.16 | 0.46 | 1.00 | | |
| 9 | 0.46 | 0.46 | 0.46 | 0.46 | 0.46 | 0.16 | 0.46 | 0.69 | 1.00 | |
| 10 | 0.54 | 0.54 | 0.54 | 0.54 | 0.54 | 0.16 | 0.54 | 0.46 | 0.46 | 1.00 |

表 3 各类内序列与其标准序列相似百分率(%)

| 类 别 | 一 | | | | | | 二 | 三 | | 四 |
|---|---|---|---|---|---|---|---|---|---|---|
| 类内标准 | 1 | | | | | | 10 | 8 | | 6 |
| 序列号 | 1 | 2 | 5 | 3 | 7 | 4 | 10 | 8 | 9 | 6 |
| 相似性(%) | 100 | 97 | 97 | 90 | 85 | 74 | 100 | 100 | 69 | 100 |

序列与其标准序列的相似百分数，结果见表 3。各标准序列间的比较则表示类间相似性百分数，结果见表 4。

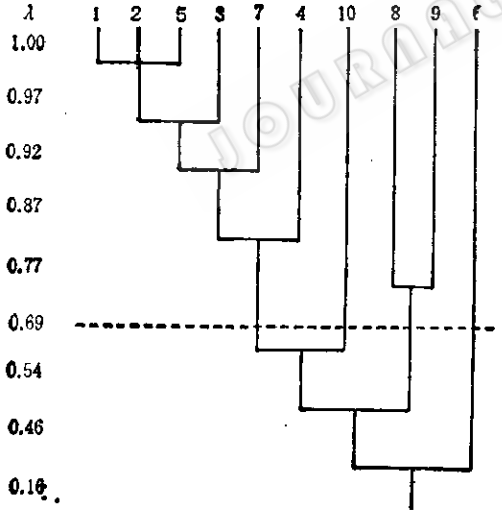


图 2 10 种包涵体病毒动态聚类图
(虚线表示划类界限)

表 4 类间相似百分率(%)

| | 一 | 二 | 三 |
|---|---|---|---|
| 二 | 51 | | |
| 三 | 46 | 26 | |
| 四 | 16 | 6 | 8 |

## (三) 讨论

1. 研究有机体进化，精确而有效的方法是直接比较某种基因或蛋白质序列。这种方法必须满足两个条件：第一，被比较的一系列基因或蛋白质必须具有相同的功能；第二，必须是单拷贝。例如细胞色素 C、组蛋白 $H_4$ 等都成功地构建过分子进化树[6]。本文所讨论的杆状病毒包涵体蛋白也满足上述两个基本条件。

2. 随着各学科间的互相渗透，模糊聚类分析方法在其它学科中的应用也越来越广泛，诸如对水质评价[7]、昆虫幼虫发育[8]、牧草产量预报[9]等都用到过这种方法。张敬宝采用这种方法对 25 种蝎毒素进行分类，与血清学结果完全吻合[10]。本文采用这种模糊聚类分析方法将上述 10 种病毒分为四类：AcMNPV, GmMNPV, OpMNPV, OpSNPV, BmSNPV, LdMNPV 等 6 种为第一类，代表鳞翅目核多角体病毒；NsSNPV 为第二类，代表膜翅目核多角体病毒；PbGV, TnGV 为第三类，代表鳞翅目颗粒体病毒；TpSNPV 为第四类，代表双翅目核多角体病毒(见图 2)。与 Rohrmann 的结果一致[2]。

为验证分类结果的准确性，对分类结果进行了类内和类间的比较。每一类内各序列间相似或同源程度高，至少不低于 69%，最高达 97%（表 3）。类间相似程度则较低，最高也只有 51%（表 4）。说明用模糊聚类分析方法的结果是可靠的。

3. 动态聚类图中 λ 值即表示相似性，对其截取是非常关键的，不同样本、不同目的须选取不同值。本文中如取 0.77 或更大，即要求属于同一类的序列同源性达 77% 以上，则类间的相似性也会高达 69%（如 PbGV 和 TnGV 间），无疑太高。反之，如取 0.54 或更小则要求太低，将导致类内相似性低至 50% 以下（如BmSNPV 和 NsSNPV 间），显然不合理。故以 $\lambda = 0.69$ 最合适。

4. 模糊聚类分析方法并不是对各序列进行简单的类别划分。就本文而言，更重要的是它还表示了同一类内各序列间同源或进化的关系及类间相似的程度。如果将动态聚类图（图 2）以序列 1 为主线，其它序列为分枝，则会“改造”成一个与 Rohrmann 的非常相似的系统进化树。它们间的很微小的差别可能是因为 Rohrmann 采用的是包涵体蛋白的全序列，而本文则是 N- 末端的39个氨基酸残基序列。

5. 本文结果表明，模糊聚类分析方法对已知序列的杆状病毒分类并构建系统进化树是切实可行的。如果对大量序列进行分类，借计算机的帮助可使工作大大简化[10]。

## 参 考 文 献


1. Maeda S: *Annu. Rev. Entomol.*, **34**: 351—372, 1989.
2. Rohrmann G F: *J. gen. virol.*, **67**: 1499—1553, 1986.
3. 胡裕文等：病毒学报，**3**(2)：156，1987。
4. 陈贻源：模糊数学，华中工学院出版社，100，1984。
5. 冯德益等：模糊数学方法与应用，地震出版社，58，1983。
6. Wilson A C et al.: *Annu. Rev. Biochem.*, **46**: 573—639, 1977。
7. 张子安等：生态学报，**7**(1)：1，1987。
8. 杨跃雄等：动物学研究，**10**(2)：88，1989。
9. 周电辉等：生态学报，**6**(1)：43，1986。
10. 张学宁等：生物物理学报，**3**(2)：141，1987。