生物工程学报 12(1):28～33,1996
Chinese Journal of Biotechnology

尝试利用 *R-Q* 对应因子分析识别分群融合子代

艾云灿* 孟繁梅 高培基 王祖农
(山东大学微生物研究所 济南 250100)

摘 要 在特定融合组合的子代群体中，以不同姊妹菌株作为观察样本（n=N），以蛋白质电泳全部谱带泳动率作为指标（p=P），则所对应的谱带光密度扫描峰面积作为其观测值（X）（当某菌株缺乏某条带时，其峰面积记为零），利用 *R-Q* 对应因子分析程序，解析这个数据矩阵（$X_{n,p}$）。能够在同一标度的因子平面上考察这一特定融合重组事件发生后，姊妹菌株间和各菌株内蛋白质谱带位置与含量间以及它们相互之间的多重内在联系，从而推断子代群体与亲本的遗传继承关系。并由此来识别和分群姊妹菌株。这种探索具有拓展数值分类能力和指导融合育种实践的意义。

关键词 *R-Q* 对应因子分析，融合子识别分群，*Aspergillus niger*，*Trichoderma reesei*，数值分类，融合育种

我们系统地改良了丝状真菌纤维素酶系细胞融合育种方法[1]。前已报道通过统计分析酶系中各组分活性表现，来判定子代群体中是否发生了真正的基因组融合重组[2]。那么如何进一步地从中迅速找出某几个可能具有典型意义的重组子呢？我们希望采用蛋白质电泳图谱的数值分析思想来实现初步搜索，这不仅是因为数值分类方法本身已有广泛的应用基础[3]，而且还因为真菌能够大量分泌包括纤维素酶系在内的胞外蛋白，使得批量操作上较简便。然后再对找出的几个典型重组子进行操作上较繁琐的 DNA 鉴定分析，才能提高针对性和有效性。

遗憾的是，70 年代发展起来的较为成熟的数值分类方法[4]自身存在两个方面的局限性，即视性状独立等同的编码标度原则和单型聚类分析（对样品 *Q* 型或对指标 *R* 型）技术，只能较好地完成属种水平上的真实分类。怎样提高属以上或种以下情形的分类有效性，是近年来不断实践探索的难题，例如谭远德等[5]提出以“（染色体）核型相似系数”为基础进行系统聚类就考虑到了性状的加权量，国外学者以高度保守的 rRNA-DNA 分子杂交为考察对象，能够拓展到某些属及属以上乃至科间的分类[6]。这些新的改进无疑都是对传统数值分类学的重要发展，但是他们都还没有彻底摆脱单型聚类这种数值分析技术本身难以满意地处理那些客观上存在极大相关性的性状问题的局限性。另一方面，针对种以下菌株间特别是在具有高度亲缘关系的融合子代群体中的姊妹菌株间，由于存在性状的复杂相关性（如样品间、指标间、以及样品与指标间都相互关联）的情形下的分类（或称识别与分群）问题，国内外都还极少见到有关这方面考虑的报道。

国家自然科学基金（No. 39400002）。中国博士后科学基金（No. 1994. 5/2154）资助课题。
* 现在通讯地址：广州市中山大学生命科学院，广州 510275。
本文于 1994 年 11 月 11 日收到。

本文介绍采用近年来发展起来的 R-Q 对应因子分析技术[7]来处理蛋白质电泳图谱信息的新尝试。结果讨论了由此克服性状间因不独立等同而彼此复杂相关所致的困境，从而改善和拓展数值分类能力的可能性以及在融合育种实践上的应用意义。

1 数学方法和生化遗传学原理

1.1 数学方法

数学上，为了简化地考察不独立或相关的多指标（多维向量）的作用，可以采用和因子分析方法，寻求少数几个潜在指标（称主因子）的线性组合来表示全部指标[7]。具体地，设有 n 个样品，每个样品测得 p 个指标的数据，则可由原始数据矩阵 $(X_{ij})_{nxp}$ ($i=1, 2, \cdots, n$; $j=1, 2, \cdots p$) 得到各指标间的相关系数矩阵 $R=(r_{ij})_{pxp}$ ($i, j=1, 2\cdots, p$)，或者得到样品间的相似系数矩阵 $Q=(q_{ij})_{nxn}$ ($i, j=1, 2, \cdots n$)。当以 R 或 Q 为出发点，进行 R 或 Q 因子分析，可以找出控制所有指标或样品的少数几个主因子，从而对指标或样品聚类；运用这种单型式的 R-或 Q-因子分析后再聚类，曾经是当年倡导数值分类的先驱者们探索过的难题[4]。

这里特别指出的是，现在数学上已经证明 R 与 Q 之间存在对应关系，通过正交变换可以从 R 型结果很容易得到 Q 型结果，其数学处理过程就是 R-Q 对应因子分析。当今高度发展了的计算机技术能够圆满地实现这种计算[7]。这样处理的结果，不仅解放了计算能力，而且由于 R 与 Q 具有相同的特征值，使得可以用相同标度的因子轴去同时表示样品和指标，从而能够科学地揭示出样品与指标间、样品之间、指标之间的多重内在联系。在 A-I 空间里同时标位 R-Q 结果的景象是先驱者们所难以想象的[4]。对计算结果的解释原则是：在同一标度的因子平面或空间里，图形上邻近的一些指标，表示它们密切相关或来自同一源；而邻近的样品点，则表示它们密切相关或属同一类群，仍是同一源的产物。其相关程度将由与它们相邻近的几种指标来表征，这就有助于对样品分类并解释其成因。详细的数学过程可参见有关专著，我们所使用的微机计算程序是参照文献[7]修改而成的。

1.2 生化遗传学依据

已知基因组融合重组后，子代群体中姊妹菌株间酶系活性表现出很强的统计规律性[2]。同时又知道，蛋白质是基因表达的直接产物具有遗传稳定性，非变性电泳图谱中包含了同工酶的信息。那么基因转录及加工、蛋白质翻译后修饰等诸多环节都可以从表观上改变蛋白质分子量大小及相对含量，反映在图谱上就是泳动率（Rf）和染色谱带扫描峰面积（As）上的差异。所以不难理解，来自于同一亲本融合组合的几个姊妹菌株间以及每个菌株内的各条蛋白质带泳动率及峰面积之间，都必然存在内在的复杂的相关性。

由此我们设想，以不同姊妹菌株作为该特定融合组合的子代群体的观察样本(N)，又以所有出现的蛋白质谱带位置（Rf）作为指标（P），则对应的谱带光密度扫描峰面积（As）作为其观测值（X_{nxp}）（当某一菌株缺乏某一条带时，峰面积记为零）。那么利用 R-Q 对应因子分析技术来解析这个数据矩阵，就能够在同一因子平面或空间里考察这一特定融合组合事件发生后姊妹菌株（样本）间、各菌株的蛋白质谱带及含量（指标）间的多重内在联系，从而推断子代群体与亲本的遗传继承关系，并由此来识别和分群姊妹菌株。

2 实例分析

作为例子，我们采用直观上就可判定结果的有限实验数据（见表1）来分析说明。

表1 蛋白质电泳图谱扫描数据

Table 1 Data of photometer-scanning of protein profiles

No.	*Rf*	A	B	C	D	E	F
1	0.12	41134.5		41101.4			
2	0.16	81974.8		91024.8			
3	0.22	71657.1		91274.1			
4	0.30	32977.1		72679.1			
5	0.33	100397.4	102148.1		117417.6	93847.1	101032.4
6	0.38	101749.3		116570.9	91295.6	86139.6	81575.2
7	0.42	148973.4	203214.8		156139.5	151295.6	151575.2
8	0.46	1089.9		1162.8			
9	0.52	1100.4		1117.1			
10	0.53				98609.0	91441.9	94125.4
11	0.58	109736.8	110796.8				
12	0.60	889.17.7		90894.3	92690.4	91425.5	91521.5
13	0.62				91214.4	91573.5	91412.7
14	0.64	79994.5	87894.3				
15	0.68	80112.2		99813.6		1163.8	1072.9
16	0.71	99713.9	100927.2	3176.4	185206.8	182360.4	187127.1
17	0.82	160128.0	187404.9	3796.8	246721.4	214694.8	212537.6
18	0.86				2813.7	2797.7	2789.4
19	0.89	89276.3		99219.2			
20	0.91					397.1	
21	0.93				713.4	701.9	717.1
22	0.96	159533.6	178328.5		169451.2	158833.5	162448.3

2.1 构建原始数据矩阵，计算获得 *R* 型和 *Q* 型因子载荷矩阵

表1中A～F共6个样品菌株，其中A是混合了亲本B（*Aspergillus niger* AMS_{11}）和亲本C（*Trichoderma reesei* QM9414）的发酵液之后的电泳图谱作对照。D～F为来自双亲本B和C融合后的三个重组子。依据 *Rf* 值不同总共有22条可辨的谱带，作为22个指标。这里某个菌株可能缺少某条带则相应的峰面积记为0。将这个数据矩阵 $(X)_{6\times22}$ 输入修改后的微机程序[7]数据库，当取临界概率 PC=0.90，计算精度 $E=10^{-7}$ 时，数秒钟内求得主因子数 M=2。这表明有90%的把握认为二个主因子 Z_1、Z_2 能够分别代表 *R*-（指标）或 *Q*-（样品）的全部信息. 相应的因子载荷矩阵输出如表2、表3。

表 2　*R*-型因子载荷矩阵

Table 2　Matrix of *R*-factor charge

Object	Z_1		Z_2		Object	Z_1		Z_2	
1	0.1747221		6.897907	E-03	12	6.811394	E-02	−0.1063925	
2	0.2612225		3.245061	E-03	13	−0.1221830		−0.1434487	
3	0.2632887		−5.324039	E-03	14	1.598466	E-03	0.2267812	
4	0.2415618		−2.987819	E-02	15	0.2721192		−5.604152	E-03
5	−9.767142	E-02	3.102075	E-02	16	−0.1417547		−5.321969	E-02
6	0.1127374		−0.1016595		17	−0.1526328		3.977891	E-03
7	−0.1269875		7.939585	E-02	18	−2.138658	E-02	−2.510572	E-02
8	2.947148	E-02	6.622239	E-04	19	0.2726757		3.668830	E-03
9	2.882417	E-02	1.017531	E-03	20	−4.609431	E-03	−5.54476	E-03
10	−0.1244089		−0.1459149		21	−1.077528	E-02	−1.264792	E-02
11	6.957883	E-03	0.2491249		22	−0.1247889		5.304417	E-02

表 3　*Q*-型因子载荷矩阵

Table 3　Matrix of *Q*-factor charge

Sample	Z_1	Z_2	Sample	Z_1	Z_2
A	0.2178440	0.1739061	D	−0.1870615	−0.1286891
B	−0.1558603	0.3114212	E	−0.1783502	−0.1318671
C	0.5813830	−0.1043343	F	−0.1816972	−0.1305508

2.2　在相同标度的同一因子平面上标位样品和指标，由点图分类并解释成因

以 Z_1 为横坐标，Z_2 为纵坐标，构成因子平面。依据表 2 值将 22 条蛋白质谱带描点其上（●），同时依据表 3 值也将 6 个菌株描点在同一因子平面上（▲）（如图 1 示）。然后参照表 1 中谱带序号，分别将属于亲本 B 和亲本 C 的全部谱带点各圈为一类。接着将 D～F 所独立拥有的谱带点也圈为一类，并从此圈的中心出发，以矢量标出存在于双亲本 B、C 圈中但却是 D～F 也具有的其他谱带点（图 1 示）。结果就一目了然了：(1) D～F 三个菌株点非常邻近，的确为姊妹菌株关系，同属一类，分布在第Ⅲ象限。而 B 和 C 分别各属一类（事实上分别是曲霉属和木霉属），分居第Ⅱ象限和第Ⅳ象限，两者简单混合（作对照）的 A 则另属一类，居第Ⅰ象限。这就精确而形象化地表示了子代之间、亲本之间、子代与亲本之间的关系；(2) 确证子代 D～F 与亲本 B 和 C 的遗传继承关系。它们都共同拥有 10、13、18、20、21 等新生的特征谱带，同时还具有亲本 B 中的 5、7、17、22 等谱带以及亲本 C 中的 6、12 等谱带。特别是 16 带虽为双亲本 B 和 C 都具有，但却在子代 D～F 中加强了（图形上更邻近），另外 A 中的特征带（11、14、19、9、8、1、2）是子代 D～F 所缺乏的。这就清楚地表明了 D～F 菌株是基因型上更接近于亲本 B（*A. niger* AMS_{11}，图形上更邻近）的重组子姊妹菌株而不是异核体或杂合二倍体（否则应与 A 同象限），更不是干扰的杂菌。显然直观结果能够被精确地表征，比目测图谱判定更科学。

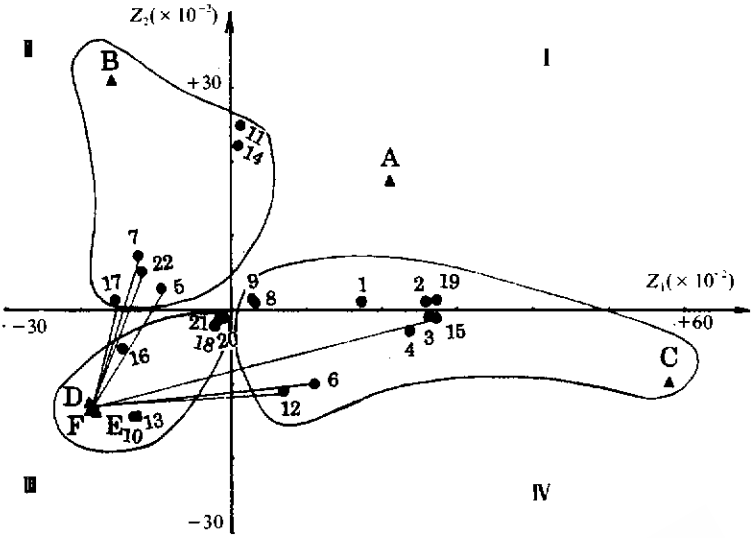

图 1 *R-Q* 对应因子分析所揭示的融合子与亲本间的关系

Fig. 1 Relationship between fusants and parants demonstrated by *R-Q* double factor analysis

Explanations were given in the text, ▲ Strains, ● Bands

3 讨论

3.1 本文仅尝试分析了一个特定实例

从菌株角度看，与对照样品（双亲简单混合）邻近且同象限，则很可能是异核体或杂合二倍体；与某个亲本邻近且同象限，则认为是非整倍体（主要保留了所邻近的那个亲本的染色体组）或者是亲本的突变株；只有当不属于上述情形而是定位分布在因子平面上剩余的那个象限且互相邻近时，才是来自同一融合组合的重组子姊妹菌株。这时可以依据其互相邻近的程度，再细分为亲缘相近程度不同的亚群。当样品（菌株）点与上述双亲本及其混合物等样品点相距甚远时，应当考虑其是干扰杂菌。另一方面，从谱带角度看，可以从已被识别分群为重组子的菌株中还存在有来自于双亲本中已存在谱带的多少和这些谱带与重组子群圈中心距的长短两个侧面来判断它们是更偏向于某亲本的重组子，从而肯定遗传继承关系，排除杂菌干扰可能性。

3.2 操作可行性问题

如同 Sokal[4] 预计的那样，实例分析也显示出，在种以下菌株水平上，单型 *R* 或 *Q* 主因子数都不会太多（一般小于 3），因此绝大多数情形都可用二维因子平面来标位。尽管如此，仍然不难想象当子代群体中姊妹菌株间变异越大时，微机计算能力越有限，在同一因子平面上标位结果就越困难。不过，旨在寻求少数典型重组子的育种目的，可以容忍我们人为地灵活干预这一过程。比如将经过酶系活性分析[2]所得到的最有希望的少数子代，分批挑取有限的几个同双亲本及其混合对照进行比较计算（当然电泳及扫描数据是可以不必经过分批就直接一次性得到的），还是能够完成识别分群的。正因如此，首先从统计意义上肯定优秀重组子的酶系活性[2]及蛋白质电泳谱带（本文）的融合重组特征，然后再深入到 DNA 水平上精细鉴别的策略和措施显得更加明智。

3.3 本法所论及的原理可能具有普遍适用性

这里谨指出，虽然本文是以蛋白质为考察对象来探讨新尝试的，但有理由相信，以其他具有类似的相关性性状为考察对象，诸如同工酶谱、DNA RFLP 图谱的解析时，也不妨可以采用这种思想方法和处理技术。而且在经过蛋白质（或同工酶）电泳图谱分析所筛选出来的有限个数重组子基础上，进一步采用类似方法，直接考察 DNA 位点多态性，可能会根本上拓展关于这些姊妹菌株间的数值分类（或识别分群）能力。笔者认为，无论是从发展数值分类方法的理论探索还是从指导融合育种实践角度看，这方面的工作无疑都是值得深入探讨的重要课题。

参 考 文 献

〔1〕艾云灿．见：中国微生物遗传学学术讨论会论文摘要集，武汉：武汉大学出版社，1994，p. 1.

〔2〕Ai Y C，Zhao X H，Yu J L. Chin J Biotechnol，1994，10 (1)：61～66.

〔3〕汪恩涛，陈文新．微生物学通报，1988，15 (6)：265～268.

〔4〕史尼斯［英］P.，索卡尔［美］R. 见：赵铁桥译，数值分类学：数值分类的原理和应用，北京：科学出版社，1984，pp. 76～77，161～162，170，180～181.

〔5〕谭远德，吴昌谋．遗传学报，1993，20 (4)：325～331.

〔6〕汪恩涛，陈文新．见：林万明主编，医学分子微生物学进展，北京：中国科学技术文献出版社，1991，pp. 383～397

〔7〕卢崇飞，高惠璇，叶文虎．环境数理统计学原理及应用程度，北京：高等教育出版社，1989，pp. 180～225.

An Attempt on Using the *R-Q* Double-factor Analysis Method to Identify and Group Fusants

Ai Yuncan　Meng Fanmei　Gao Peiji　Wang Zunong

(*Institute of Microbiology, Shandong University, Jinan* 250100)

Abstract　Among the fusants from the definite fusion-cross, the diffrent sister-stains were regarded as the samples of observation ($n=N$), and the positions of total bands of protein profiles as the objects ($p=P$). The photometer-scanning area of the specific band as the experimental value (X) (Zero was taken when the specific band of certain strain was absent). The genetical multi-relationships among the inter-and intra sister-strains in terms of the position and content of protein bands after the fusion recombination occurred from this definite fusion-cross couid thus be determined on the same orientational factor-plate by using the computer program of *R-Q* double-factor method to analysis this data matrix $(X)_{nxp}$. Therefore the sister-strains could be identified and grouped from the deduced hereditysship between the fusants and parents.

Key words　*R-Q* double-factor analysis, fusants identifying and grouping, *Aspergillus niger*, *Trichderma reesei*, numberical taxonomy, cell-fusion breeding