第26卷 第2期 作 物 学 报 Vol 26, No. 2
2000 年3月 ACTA AGRONOMICA SINICA Mar., 2000

构建作物种质资源核心库的一种有效抽样方法*

徐海明 胡 晋 朱 军

(浙江大学农学系, 浙江杭州, 310029)

提 要 本文提出了基于基因型值构建作物种质资源核心库的抽样方法。采用包括基因型与环境互作的遗传模型及混合线性模型统计分析原理, 无偏预测基因型值。用基因型值计算基因型间的马氏距离, 并采用不加权类平均法进行聚类。根据树型图, 确定合理的分类水平, 将群体分成若干不同的类群。计算各基因型的平均离差度, 在各类群内按确定的比率, 选取平均离差度大的材料构建资源核心库。以棉花168个基因型5个纤维性状为例构建核心库, 所获得的48个核心资源能保存原棉花资源的遗传多样性。

关键词 统计抽样方法; 基因型值; 核心资源; 种质资源

An Efficient Method of Sampling Core Collection from Crop Germplasm

XU Hai-Ming HU Jin ZHU Jun

(*Department of Agronomy, Zhejiang University, Hangzhou*, 310029)

Abstract A method for sampling core collection from crop germplasm was proposed. A genetic model with GE interaction and mixed model approaches were used for analyzing the genetic data. Mahalanobis distance among varieties calculated from predicted genotype values were employed for clustering crop germplasm using unweighted pair group method with arithmetic average of hierarchical cluster. After specifying the appropriate threshold value of classification based on the dendrogram, all genotypes could be clustered into some different sets. The mean of deflection for each genotype was calculated. The core collection was constructed by the core entries sampled from each set, with the larger mean of deflection. A worked example on 168 varieties of cotton with five fiber traits was presented. The results showed that the 48 genotypes as a core collections could represent the genetic diversity of original resources.

Key words Statistical sampling method; Genotypic value; Core collection; Germplasm

Franke 于1984年最早提出核心资源(core collection)的概念[1], 认为核心资源是以最少数量的遗传资源包含一个作物种及其近缘种的最大限度的遗传多样性。资源核心库应包含需优先保存的种质材料, 而其余的基因型可以作为备用资源库。建立资源核心库主要是为对其进行优先评价和利用, 从而提高整个种质库的管理和利用水平。近年来, 核心资源的研究与发展日益受到重视, 国外已构建了一些种质资源的核心库[2~4]。

* 国家自然科学基金资助项目(39470377)
致谢 辽宁省经济作物研究所李瑞祥、王存晋研究员为本文提供研究所需的实例、数据, 特致谢意。
收稿日期: 1999-01-28, 接收日期: 1999-05-19

构建种质资源核心库的通常方法是: 根据已有的资料, 用一种分组方法, 将相似的种质资源归为一组, 在已鉴定的组内, 可根据一定的原则或实际情况来选取需优先保存的核心资源。建立的核心库必须尽可能多地保存原有种质资源的遗传变异, 因此, 选取核心资源的策略至关重要。Peeters 等[5]指出, 聚类分析可使核心资源这一概念成为可能。胡晋等[6]提出了用多次聚类的方法构建资源核心库, 这一方法能够保存原有资源库的遗传变异的模式, 尽管选取的核心资源也能够保存原有的遗传变异, 但不能最大程度地保存原有资源群体的遗传变异量。本文提出了一种选取核心资源的新方法, 在保存原有资源群体的遗传变异模式的前提下, 能使建立的核心库最大可能地保存原有群体的遗传变异量。

1 模型和方法

1.1 基因型效应值的预测 农业试验中存在大量试验误差, 作物性状又多与环境存在着互作, 直接用表型值所得的分类结果不能正确反映资源群体固有的遗传结构[7]。种质资源遗传试验一般按田间行列编号顺序种值基因型, 以一定间隔穿插对照基因型, 控制田间不同位置的差异。对于多年份的这类遗传试验, 一般存在环境效应、环境内的行效应、环境内的列效应、基因型效应、基因型与环境互作效应, 以及机误[8, 9]。因此得到的观察值可用混合模型作如下分解:

$$Y_{hg(ij)} = \mu + E_h + R_{i(h)} + C_{j(h)} + G_{g(ij)} + GE_{hg(ij)} + e_{hg(ij)}$$

其中 E_h 表示第 h 个环境的效应, 固定效应;

$R_{i(h)}$ 表示环境 h 内第 i 行的效应, 固定效应;

$C_{j(h)}$ 表示环境 h 内第 j 列的效应, 固定效应;

$G_{g(ij)}$ 表示在环境 h 内第 i 行 j 列处的第 g 个基因型效应, 随机效应, $G_{g(ij)} \sim (0, \sigma_G^2)$;

$GE_{hg(ij)}$ 表示环境 h 与基因型 g 的互作效应, 随机效应, $GE_{hg(ij)} \sim (0, \sigma_{GE}^2)$;

$e_{hg(ij)}$ 是机误效应, 随机效应, $e_{hg(ij)} \sim (0, \sigma_e^2)$。

采用朱军[10~12]的混合线性模型统计分析方法进行统计分析, 无偏预测基因型效应值。

1.2 遗传距离的计算 因为性状间存在相关性, 故本文采用马氏(Mahalanobis)距离计算方法计算基因型间的遗传距离。假设共有 n 个基因型, 采用 m 个性状进行聚类。第 i 个基因型与第 j 个基因型的基因型效应向量分别为 $g_i^T = (g_{i1}, g_{i2}, \cdots, g_{im})$, $g_j^T = (g_{j1}, g_{j2}, \cdots, g_{jm})$, 则第 i 个基因型与第 j 个基因型间的马氏距离计算公式为[13]

$$D_{ij}^2 = (g_i - g_j)^T (V_G)^{-1} (g_i - g_j)$$

其中 $V_G = (\sigma_{ij})$, $i = 1, 2, \cdots, m$; $j = 1, 2, \cdots, m$。

1.3 聚类 采用系统聚类法中的不加权类平均法(UPGMA)。假设类 G_i 与 G_j 分别有 n_i 与 n_j 个基因型, 其合并所得的新类记为 G_r, 有 $n_r (= n_i + n_j)$ 个基因型, 它与其它各类 G_s 的类间距离计算公式为[14]。

$$D_{rs}^2 = \frac{n_i}{n_r} D_{si}^2 + \frac{n_j}{n_r} D_{sj}^2$$

1.4 构建资源核心库 可按种质库资源总数的一定比率(p), 从某一作物的种质库中选取适量的材料构成核心库。根据对保持群体遗传变异量或等位基因变异的要求, 选取资源库总量的 20% ~ 30% ($p = 0.2 \sim 0.3$) 作为样本, 被认为具有较好的代表性[15]。

采用马氏距离对所有基因型进行系统聚类后, 借助得到的树型图, 确定合理的分类水

平, 可对基因型进行分类, 从而得到若干不同的类群。对类群内的各基因型分别计算相对于群体的平均离差度, 计算公式为,

$$s_i^2 = \frac{1}{m}\sum_{j=1}^{m}\frac{g_{ij}^2}{\sigma_j^2},\quad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m$$

其中 σ_j^2 是群体的第 j 个性状的基因型方差。

根据分类结果及各类容量的大小, 按所确定的比率($p = 0.2 \sim 0.3$) 计算在各类群内抽取核心材料的数目。容量较小(抽取数目不到 1) 的类群, 其抽样数可定为 1。在各类群内, 定量选取具有较大平均离差度的材料, 构成资源核心库。

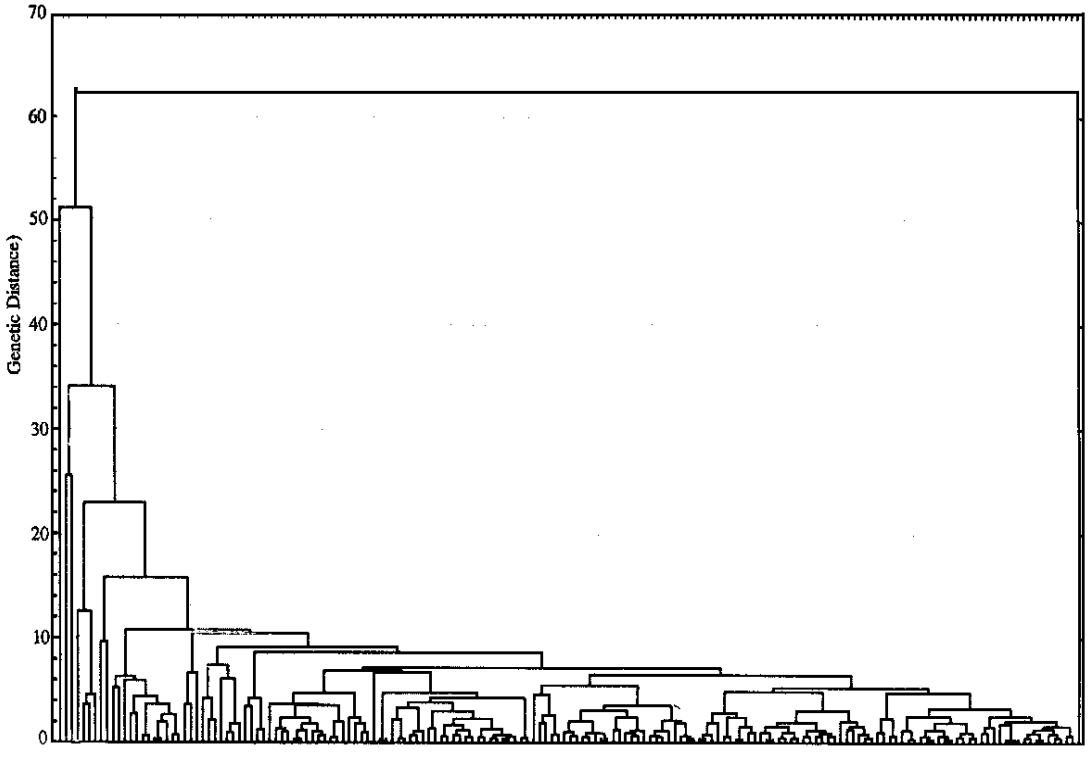

图1　棉花5个纤维性状168个基因型的聚类树型图

Fig. 1　Dendrogram of cluster of 168 genotypes based on five fiber traits of cotton

2　应用实例

现以辽宁经济作物研究所的棉花种质资源168个基因型, 5个纤维性状(2.5% 跨长(mm)、整齐度(%)、强度(gf/tex)、伸长度(%)、麦克隆) 为例, 构建核心库。首先采用朱军[10]提出的调整无偏预测(AUP) 法无偏预测遗传效应值, 用所得的预测值计算基因型间的马氏距离, 采用系统聚类法中的不加权类平均法进行聚类, 根据树状图确定遗传距离为6.7时进行分类, 共得17个不同的类群。

计算各基因型的平均离差度, 每类内分别选取25% 的基因型作为核心资源。如表1列出了其中一个类群的基因型预测值及平均离差度。这一类共有12个基因型, 需选取3个作为核心资源。选取平均离差度最大的3个基因型(141, 143, 42) 作为核心资源。

表1 一个类群的基因型预测值及平均离差度

Table 1 The predicted genotypic values of traits and mean of deflection on one cluster

基因型 Genotype	2.5%跨长 Length(2.5%)	整齐度 Uniformity	强度 Strength	伸长率 Elongation	麦克隆 Micronaire	平均离差度 Mean of deflection
141*	-1.283	2.776	2.956	0.980	0.385	4.266
143*	-1.077	3.404	2.401	0.906	0.156	3.528
42*	-0.692	1.286	2.357	0.881	0.480	2.831
102	-2.924	3.111	1.032	0.694	0.222	2.702
19	-2.370	2.098	1.652	0.351	-0.109	1.747
62	-1.108	2.205	1.743	0.566	0.214	1.740
5	-1.104	1.408	1.630	0.643	0.340	1.630
51	0.168	1.765	1.641	0.494	0.270	1.348
26	0.150	0.409	2.093	0.573	-0.006	1.340
140	-0.985	1.711	1.545	0.407	0.116	1.128
8	-1.346	0.692	1.218	0.368	0.338	0.962
48	0.477	0.962	1.276	0.445	0.198	0.784
168个基因型的方差	2.075	1.963	0.926	0.176	0.087	

*: 抽取作为核心资源的基因型。Genotypes sampled for core collection

按这一抽样标准分别对每一类群进行抽样，共得48个基因型，构成资源核心库。基因型代号分别为：1, 167, 55, 127, 93, 87, 88, 38, 130, 92, 169, 47, 139, 141, 143, 42, 75, 118, 149, 31, 11, 112, 7, 17, 12, 170, 15, 103, 101, 114, 145, 59, 61, 60, 123, 104, 84, 36, 161, 109, 168, 77, 29, 21, 146, 82, 144, 111。

3 资源核心库的评价

核心资源应能代表原有种质资源群体的遗传多样性。构建的核心库是否能很好地代表原种质资源群体的遗传多样性，可用方差、极差、均值和变异系数等参数评价。核心库各性状

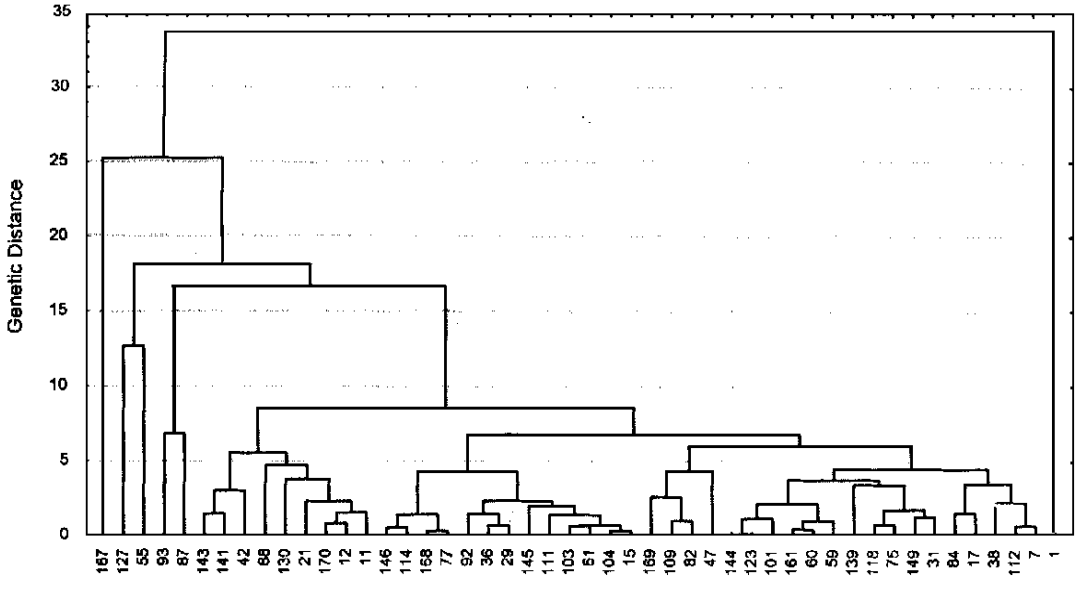

图2 棉花5个纤维性状48个基因型的聚类树型图

Fig. 2 Dendrogram of cluster of 48 genotypes based on five fiber traits of cotton

的方差和变异系数应不小于原群体的方差和变异系数, 而极差与均值则应基本保持不变。

对棉花5个纤维性状的基因型值构建的核心资源进行评价。用 F 统计检验检测资源核心库与原群体间方差的显著性差异, 在此基础上, 对均值间的差异显著性进行 t 检验。结果表明核心库与原群体之间五个性状的方差都存在极显著的差异(0.01显著性水平), 而均值都无显著性差异(表2)。对于极差, 只有麦克隆为1.908, 略小于原样本的极差2.045, 而其余四个性状的极差则与原样本相同。从变异系数看, 核心资源的变异系数比原样本的变异系数明显增大。因此, 可以认为采用本文所介绍的方法构建的资源核心库, 能以较小的样本容量保持更大的遗传变异。

表2　棉花核心资源与原资源样本5个纤维性状的比较

Table 2　Comparison between core collections and original accessions of cotton on five fiber traits

性状 Trait		方差 Variance	均值 Mean	极差 Range	变异系数 C.V.
2.5%跨长	168个基因型	2.075	26.504	10.198	0.054
Length(2.5%)	48个基因型	4.867**	26.307	10.198	0.084
整齐度	168个基因型	1.963	50.720	9.924	0.028
Unifommity	48个基因型	4.477**	50.800	9.924	0.042
强度	168个基因型	0.926	20.011	7.878	0.048
Strength	48个基因型	2.072**	19.869	7.878	0.072
伸长度	168个基因型	0.176	6.064	2.794	0.069
Elongation	48个基因型	0.375**	6.042	2.794	0.101
麦克隆	168个基因型	0.087	4.053	2.045	0.073
Micronaire	48个基因型	0.169**	4.037	1.908	0.102

** 48个基因型的方差与所有基因型间差异达0.01显著水准。

** Significant at 0.01 level between variances of all genotypes and 48 genotypes

4　讨论

种质资源核心库由种质库中的一部分材料所组成, 用于代表整个种质库的遗传范围[16]。IBPGR (国际植物遗传资源委员会)认为, 核心库能以最少的重复代表一个种及其野生近缘种的遗传多样性[17]。核心库中入选的材料都具有代表性, 互相之间都存在着生态上或遗传上的距离。为了使核心库能用尽可能少的基因型保存尽可能多的遗传变异, 同时又能反映整个种质资源群体的遗传结构, 应从各类群内根据容量的大小选取一定比列的核心资源, 入选的材料应具有较大的平均离差度。

聚类分析作为一种重要的研究工具已被广泛应用于种质资源的分类、基因型遗传差异的评价等研究中[5, 18]。为使构建的资源核心库能保存原有群体的遗传结构, 首先要对该群体进行遗传分类, 确定不同材料之间的遗传关系, 再从不同的类群中选取一定数目的基因型作为核心资源。

在目前的聚类分析中, 大多直接利用表现型值进行分类, 构建核心资源所用的数据大多为表现型值。由于作物的重要经济性状多数为数量性状, 并存在基因型与环境的互作, 同时遗传试验又存在试验误差, 用表现型值计算的遗传距离不能正确度量基因型间的遗传差异, 所得的遗传分类不能真实地反映种质资源固有的遗传结构[7]。因此, 种质资源的遗传聚类首先必须用合理的统计模型及统计分析方法, 排除农业试验中存在的试验误差、环境效应、基

因型与环境的互作效应，然后用无偏预测的基因型值进行遗传聚类。

在聚类分析中，分类水平的确定，目前主要借助于所得的树状图及分类基因型的专业知识。完成基因型分类后，需要对各类群进行抽样，抽样的技术直接关系到核心库的好坏。目前，常采用每组随机取同样数量的材料，或以组内材料数量按比例随机取样来构建资源核心库。前者未考虑到分组的大小，后者则未能顾及各组遗传变异的关系及差异。确定各类群的抽样数目以及抽样方法还有待于深入研究。本文根据分组的大小按比率确定抽样数目，按平均离差度抽取基因型作为核心材料，构建资源核心库。研究结果表明，用这种抽样方法构建的资源核心库能最大限度地保存遗传变异。

构建的资源核心库是否有效地保存了原有资源群体的遗传变异，可以用各性状的方差、均值、变异幅度(极差)、变异系数等参数评价。Diwan[19]认为，若满足以下条件：(1) 少于30%的性状，它们的均值及变异幅度与原资源群体的均值与变异幅度存在显著性差异(α=0.05)，(2) 对于各性状，资源核心库与原群体的变异幅度之比不低于70%，则可认为此核心库代表了原有资源群体的遗传变异。棉花5个纤维性状的研究结果表明，48个基因型与原所有基因型之间的方差有显著性差异，各性状的变异系数都有较大的提高，而各均值都不存在显著差异。对于极差，除麦克隆为1.908，略小于原样本的极差2.045外，其余四个性状都保存了原群体的变异幅度。因此，可以认为构建的资源核心库基本保持了原群体的遗传变异。本文提出的按平均离差度大小取样的方法可行。

参 考 文 献

1 Frankel O H. In: Arber W *et al*. *Genetic Manipulation: Impact on Man and Society*. Cambridge: Cambridge University Press, 1984. 161～170

2 Brown A H D, J P Grace, S S Speer. Glycine. *Soybean Genet Newsl*, 1987. 14, 9～17

3 Holbrook C C, W. F. Anderson, R. N. Pittman. *Crop Sci*., 1993, 33: 859～861

4 Diwan N, G R Bauchan, M S Mclntosh. *Crop Sci*., 1994, 34: 279～285

5 Peeters J P, J A Martinelli. *Theor. Appl. Genet*., 1989, 78: 42～48

6 胡晋，徐海明，朱军. 生物数学学报，1999(in press)

7 Steven D Tanksley, R McCouch Susan. *Science*, 1997, 277: 1063～1066

8 朱军. 浙江农业大学学报，1994, 20: 551～559

9 朱军. 遗传学报，1996, 23: 53～58

10 朱军. 生物数学学报，1993, 8(1): 32～44

11 Zhu J. Ph. D. Dissertation, *North Carolina State Univ. Raleigh, N. C.* 1989

12 朱军. 生物数学学报，1992, 7(1): 1～11

13 Mahalanobis P C. *Proc Natl Inst Sci India*, 1936, 2: 49～55

14 Sokal R R, C D Michener. *Univ Kansas Sci Bull*, 1958, 38: 1409～1438

15 Yonezawa K, T Nomura, H Morishima. In: T Hodgkin et al. *Core Collection of Plant Genetic Resources*. Chichester: Published by John Wiley & Sons, 1995. 35～53

16 Brown A H D. *Genome*, 1989, 31: 818～824

17 IBPGR. Annual Report 1990. *International Board for Plant Genetic Resources*, Rome, 1991

18 Hintum van T J L. In: Hodgkin, T, et al. *Core Collections of Plant Genetic Resources*. Chichester: John Wiley & Sons, 1995. 23～34

19 Diwan N, M S Mclntosh, G R Bauchan. *Theor Appl Genet*, 1995, 90: 755～761