浙 江 林 学 院 学 报 2007, **24**(1): 105—109
Journal of Zhejiang Forestry College

文章编号: 1000-5692(2007)01-0105-05

基于克隆文库筛选的池设计

管 宇
(浙江林学院 理学院, 浙江 临安 311300)

摘要: 池设计(pooling design)是基于克隆文库筛选(clone library screening)的一种实验设计, 要求从大量的克隆体中筛选出其中含有某组特定核苷酸串的克隆体, 它属于组合群试。一个池设计方案用一个二元关联矩阵(称为分离矩阵)表示, 行对应于试验, 列对应于克隆集合的体; 考虑到生物实验存在误差, 分离矩阵还需要有查错和纠错能力。利用子集或子空间之间的包含关系可以构建出分离矩阵, 它的每行和每列中元素 1 的个数分别都一样, 但试验总次数即行数明显大于信息下界。参 13

关键词: 克隆文库; 池设计; 群试; 分离矩阵; 信息界
中图分类号: O157.2; S11^{+}9 文献标识码: A

在 DNA 克隆文库建成后, 生物学家希望重排克隆, 即沿着 DNA 分子重建克隆的相对位置。其思路是使用一个易于确定的指纹图谱来标记每个克隆, 而指纹图谱则可看作一套出现在一个克隆中的"关键词"。如果 2 个克隆基本上交叠(overlapping), 那么它们的指纹图谱应该相似的; 而不交叠的克隆不可能有相似的指纹图谱, 于是这些图谱可以使生物学家区分交叠和非交叠的克隆, 并重构克隆的顺序。探针(probe)可以是短的随机序列或是任何在实验中已鉴定的 DNA 片段, 一个特别有用的探针类型是序列标签位置(sequence tag site, STS)。STS 技术导致具有唯一探针的作图, 并因此构建出第 1 张人类基因组的物理图谱[1]。如果探针是短的随机序列, 那么它们可能与 DNA 在很多位置杂交, 于是导致非唯一探针的作图。池设计正是由此产生的一种实验设计[2—4], 它属于组合群试。文章首先介绍群试(group testing)和池设计的基本思想, 然后重点讨论非自适应(non-adaptive)池设计及其信息界, 指出 Macula 法和 Ngo-Du 法的分离矩阵设计远没有达到其信息下界。

1 群试

1.1 群试概念

设总体容量为 N, 其中有特征 P 的个体有 d 个(d 可以已知也可以未知), 要求设计一个总次数尽可能少的试验, 找出全部特征 P 的个体, 这就是群试。

譬如要检验 150 人中哪些人携带病毒, 如果一个一个地检验则是件非常费时费力的工作。我们可以这样来操作: 先将 150 人平分为 10 组, 每组 15 人; 每个人的血样都分成 2 份, 一份留做备用, 另一份和小组中其他 14 人的血样合在一起组成一个大瓶; 先检验 10 只大瓶的血样, 如果某只大瓶不带

收稿日期: 2006-02-18; 修回日期: 2006-05-22
基金项目: 浙江省自然科学基金资助项目(Y104420); 浙江省教育厅资助项目(20040507)
作者简介: 管宇, 副教授, 从事统计计算和生物数学等研究。E-mail: guanyu@zjfc.edu.cn

病菌则对应小组的 15 个人都不带病菌；如果某大瓶带病菌，则对应小组的 15 个人中至少有 1 人带病菌，再分别检验这 15 个人各自的备用血样。如果已知 150 人中有 5 人携带有病菌，则最多时(5 人分布在 5 个不同小组)需要 10＋5× 15＝85 次化验，最少时(5 人恰在同一小组)只需要 10＋15＝25 次化验。显然病菌携带率越低，群试所需总检验次数越低，试验时间和费用越节省。当然方案有很多种，我们需要的是其中总次数最少或某种约束条件下的最少。群试开始于 20 世纪上半叶第二次世界大战中的体检工作，现在已被广泛使用于低发生率的试验设计中，如医学、生物学、军事和质量检验等[5]。

1.2 群试分类

按总体中具有某种特征 P 的个体的个数 d 是明确的数字还是概率，群试分为组合群试和概率群试。前者如：一堆产品中不慎混入若干只次品(个数已知)，要用尽可能少的试验次数找出全部次品。后者如：现要对某村所有人检查某种病，具体有多少人患病预先是不知道的，但根据经验或以往数据可估计患病人的比率。次品的个数和发病率的大小显然是影响群试的主要参数。

若后一次试验方案是根据前面试验结果进行的称之为自适的(adaptive)，如前面举的验血例子；如果所有各次试验方案彼此独立，不管前面试验结果如何，后面的试验都按预先设定的方案进行，则称为非自适的(non-adaptive)。自适算法所需试验总次数不多于非自适算法，但由于前因后果关系自适算法必须分批分阶段进行，而非自适算法可以同时进行所有试验(如果条件允许的话)，而且在试验时不必考虑各次试验之间的影响，在只要会用而不必知道所以然的实用操作软件盛行的今天，非自适应算法独占鳌头。文章主要介绍非自适应算法。

2 池设计

2.1 池设计概念

设有一克隆文库(总体)，其中有些克隆体(个体)含有某组特定的核苷酸串(特征 P)，这些个体表征为 1 或称为正(对应试验结果)，其他的个体表征为 0 或称为负。从克隆文库中取出若干个克隆体放置在一起称为一个池(pool)。一个池检验(pool assay)是一个决定这个池中是否至少有一个克隆体含有特定的核苷酸串(简称为正)的检验：将能与特定的核苷酸串杂交的探针置入池中，如果发生杂交现象则说明池中有特定的核苷酸串，即池中至少有一个克隆体(但无法知道到底有几个及哪几个是)含有特定的核苷酸串(简称为正)；如果没有杂交现象则说明池中所有克隆体都不含有特定的核苷酸串(简称为负)。取自总体的池的集类称之为一个池设计。

2.2 矩阵表示

设总体容量为 N, 其中至多有 d 个是正的。一个池设计能用一个二元关联矩阵 M 表示: 每一列代表一个个体(克隆体), 每一行代表一次试验(池); M 的元素取值 0 或 1。$m_{ij}=1$ 表示第 j 个体参加第 i 次试验(第 j 克隆体在第 i 池中), $m_{ij}=0$ 则代表第 j 个体没有参加第 i 次试验(第 j 克隆体不在第 i 池中)。

设 M 有 T 行, 相当于总共做 T 次试验, 则 T 次试验的结果可表示成 T 维向量 $V=(v_1, \cdots, v_T)^T$, 其中第 i 次试验结果为正时, $v_i=1$, 否则 $v_i=0$。注意: V 实际上就是 M 中具有特征 P 的列的布尔和(即 0＋0＝0, 0＋1＝1＋0＝1＋1＝1。以下不特别说明相加都求布尔和)。另外还可将 M 的列 C 看作基集 $\{1, 2, \cdots, N\}$ 的子集 S: 当 C 中行 i 值为 1 时, $i \in S$。

2.3 d-分离矩阵(d-disjunct matrix)

所谓列 c_0 被 k 个列 $c_1, \cdots, c_k$ 覆盖(cover), 如果 $c_0 \leqslant c_1+\cdots+c_k$。称 M 是 d-分离矩阵, 如果任一列都不能被另外 d 列覆盖。一个 d-分离矩阵能够检验出不超过 d 个正克隆, 一个克隆是正的充要条件是它作为列包含在输出结果 V 中, 即该列被 V 覆盖。

例 1 的 M 是 6×4 的 0-1 矩阵, 4 列分别对应 4 个克隆体, 共做 6 次试验(池)。第 1 次试验时池中只放入克隆体 1 和 2, 第 2 次试验时池中只放入克隆体 1 和 3, 其他类似。假如克隆体 3 含有特征 P 而其他克隆体不含有特征 P, 6 次试验后必然是第 2 次, 第 4 次和第 6 次试验结果为正, 第 1 次, 第 4 次和第 5 次

例 1[5]　$M=\begin{pmatrix}1&1&0&0\\1&0&1&0\\0&1&1&0\\1&0&0&1\\0&1&0&1\\0&0&1&1\end{pmatrix}$。

为负。假如克隆 1 和 3 含有特征 P 而克隆体 2 和 4 不含有特征 P，必然是第 1 次，第 2 次，第 3 次，第 4 次和第 6 次试验结果是正，第 5 次是负，即为 M 的第 1 列与第 3 列相加对应结果。

容易验证 M 是 2- 分离矩阵，即根据 6 次试验结果可以完全清楚哪一个或哪 2 个是正克隆体。假如 6 次试验结果全为正，意味着至少有 3 个克隆体是正的，因为任意 2 列的布尔和都不可能是分量全为 1 的向量，至于哪 3 个是甚至 4 个可能都是正的则不得而知。因此，对于多于 d 个正克隆的检验，d- 分离矩阵无能为力。

2.4　纠错能力（error-correcting capability）

在生物学特别是分子生物学实验中，产生误差以致导出错误的结论往往无法避免。Macula[6] 提出 d-分离矩阵的纠错能力概念 d^e-分离（d^e-disjunct）。一个 d- 分离矩阵称为 d^e- 分离，如果对于任意 $d+1$ 列，存在一列，它的元素中有 $e+1$ 个 1（其他元素可以是 0 也可以是 1）不能被其他 d 列覆盖（即其他 d 列对应的 $e+1$ 元素的值都是 0）。

显然，d^e- 分离矩阵的任意 2 列的 Hamming 距离至少为 $e+1$，因此可以用来查错和纠错。可以证明 d^e- 分离矩阵能够检查出 $e-1$ 个错误和纠正 $[(e-1)/2]$ 个纠错[7]。

例 2　分析例 1 的 M。它是 2^0- 分离矩阵和 1^1- 分离矩阵。即若用它来做查找不超过 2 个正克隆的池设计，必须没有实验误差；如果用它来做查找只有一个正克隆的池设计，可以容许有一次实验误差。譬如 6 次试验结果是 {1, 1, 1, 0, 0, 1}（1 代表正，0 代表负），如果没有实验误差则无论如何假设都根本达不到这样结果的；若容许有一次错误，则可看作 {0, 1, 1, 0, 0, 1} 加上第 1 次试验错误将 0 误变成 1，如此则得出克隆体 3 是正的唯一结论（其他如认为克隆 1 是正的之类结论都意味着至少有 2 次试验错误）。

2.5　构建 d^e-分离矩阵

现今 d^e- 分离矩阵已成为池设计的最重要工具。一个 d^e- 分离矩阵对应于一个池设计。构建 d^e- 分离矩阵的方法主要有：Macula[6,8] 利用有限集的子集之间的包含关系和 Ngo 等[9] 利用有限域上射影空间的子空间来构造。

Macula 法　设 $M(m, k, d)$ 是 $C_m^d \times C_m^k$ 的 0, 1 矩阵（$d < k < m$）：集合 $\{1, 2, \cdots, m\}$ 中任取 d 个元素对应于行，任取 k 个元素对应于列，如果 d 个元素集（行）是 k 个元素集（列）的子集，则对应的 $M(d, k, m)$ 中元素值取 1，否则为 0。Macula 证明了[8] $M(m, k, d)$ 是 d- 分离矩阵，并给出了变成 d^e- 分离矩阵的方法[6]。

当总体容量 N 很大时，往往 d 也大，那么试验次数 C_m^d 因为 d 的大而变得太大。Macula[10] 提出对于行取 $d = 2$ 而忽略其他值，当然这样就不能保证所有的正克隆都能被检查出；但当 d 与 2 相差不大时，以较大概率保证结果。

Ngo 等[9] 利用有限域上射影空间的子空间包含关系构造出 d^e-分离矩阵，并讨论了查错和纠错能力。当 Macula 法中的集的基数是素数或素数幂时，Ngo-Du 法是 Macula 法的优化。实际上，Macula 法的集合的全部子集之间的包含关系（一种偏序关系）完全可看作定义在有限域上的普通欧氏空间的子空间之间包含关系，而欧氏空间中 2 个成比例的点在射影空间上往往变成一个点，因此 Ngo-Du 法是集合的部分子集之间的包含关系（一种偏序关系）。

2.6　小 d 值的 d^e-分离矩阵

Yachkov 等[7] 进一步研究 Ngo-Du 法，得到“some happy surprise”结果。

考虑 m 维（射影）空间，向量元素取自有限域 $G_F(q)$，其中 q 是素数或素数幂。k 维子空间的个数为：

$$\begin{bmatrix} m \\ k \end{bmatrix}_q = \frac{(q^m-1)(q^{m-1}-1)\cdots(q^{m-k+1}-1)}{(q^k-1)(q^{k-1}-1)\cdots(q-1)}。 \tag{1}$$

由 Ngo-Du 法所得的 $M(m, k, r)$ 是相应于取自 m 维空间所有 r 维子空间作为行及所有 k 维子空间作为列的 0-1 矩阵。

定理 1[7] 设 $k-r\geqslant 2$, 记 $p=\frac{q(q^{k-1}-1)}{q^{k-r}-1}$, $0<r<k<q$, 则对于 $1\leqslant d\leqslant p$, $M(m, k, r)$ 是 d^e-分离的, 其中:

$$e=q^{k-r}\begin{bmatrix} k-1 \\ r-1 \end{bmatrix}_q-(d-1)q^{k-r-1}\begin{bmatrix} k-2 \\ r-1 \end{bmatrix}_q。 \tag{2}$$

特别取 $d=q^r$ 时,

$$e=\begin{bmatrix} k-1 \\ r-1 \end{bmatrix}_q+(q^r-1)\begin{bmatrix} k-2 \\ r \end{bmatrix}_q。 \tag{3}$$

由此可知, 如果 r 取 1 则 $d=q$ 值就能取到小的数。

例 3[7] 设 $q=5$, 则 $M(8, 4, 1)$ 是 97 656× 200 525 284 806 的 5^{25}-分离矩阵。换言之, 鉴别来自 2×10^{11} 个克隆的 5 个正的, 至多需要 10^5 个试验池和 25 个错误被容许。

3 群试的信息界

设总体容量为 N, 其中只有一个个体具有特征 P, 记总的试验次数为 T。每次试验有两种可能结果: 正或负, 则 T 次试验共有 2^T 种可能结果; 由于 N 中任何一个都可能或不可能具有特征 P, 即有 $2N$ 种可能; 因此要想找出特征 P 个体, 必须 $2^T\geqslant 2N$, 即试验次数 T 至少要 $[\log_2 N]+1$(这种最小值常称为信息下界)。群试的信息下界可能达到也可能达不到[11], 好的群试方案自然应该等于或近似等于信息下界。

对于总体容量为 N, 其中有不超过 d 个个体具有特征 P 的一般情形, 总试验次数 T 的信息下界 T^* 满足不等式(当 N 和 d 充分大时)[12]:

$$\frac{d^2}{2\log_2 d}[1+o(1)]\log_2 N\leqslant T^*\leqslant d^2\log_2 e[1+o(1)]\log_2 N。 \tag{4}$$

下面估算定理 1 与不等式(4)的吻合程度。不妨直接设 $N=\begin{bmatrix} m \\ k \end{bmatrix}_q$, $T=\begin{bmatrix} m \\ r \end{bmatrix}_q$, e 满足式(3)。当 m 充分大:

$$N=\begin{bmatrix} m \\ k \end{bmatrix}_q=\prod_{i=1}^{k}\frac{q^{m+1-i}-1}{q^i-1}=\prod_{i=1}^{k}[(q^{(m+1-i)/i})^{i-1}+(q^{(m+1-i)/i})^{i-2}+\cdots+1]。$$

$$=\exp\left\{(\ln q)\sum_{i=1}^{k}(m+1-i)(i-1)/i\right\}+W \qquad (W \text{ 是比前面式子低价的项})$$

$$=q^{[m-(k-3)/2]k}(1+o(1))。$$

类似 $T=\begin{bmatrix} m \\ r \end{bmatrix}_q=q^{[m-(r-3)/2]r}(1+o(1))$, $e=q^{[k-1-(r-3)/2]r}[1+o(1)]$。则:

$$d^2\log_2 e[1+o(1)]\log_2 N=(d\log_2 q)^2[m-(k-3)/2][k-(r-1)/2]kr[1+o(1)]\ll T。$$

因此, Ngo-Du 法 d^e-分离矩阵设计远没有达到其信息下界, Macula 法当然也是如此。它们每行每列出现 1 的个数都是分别相等的, 即参加每次试验的个体数目都一样, 每个个体参加试验的总次数也分别相等, 这正是这 2 种方法的总试验次数往往很大的原因所在。要想降低试验次数, 只有 d^e-分离矩阵的行列元素 1 的数目不都一样, 譬如控制在一定区间中波动, 这方面工作的难度相对比较大, 还未见到理想的结果。当然也可以分成 2 个阶段试验, 相当于将自适应算法与非自适应算法相结合, 甚至引进概率的方法来降低试验次数[13]。

随着生物学研究的进一步深入, 新的实验问题的提出, 包括池设计在内的非常年轻的计算分子生物学(computational molecular biology)越来越成为分子生物学不可缺少的组成部分, 还有太多的工作正

等待着人类去完成。

参考文献：

[1] FOOTE S, VOLLRATH A, HILTON A, *et al*. The human Y chromosome: overlapping DNA clones spanning the euchromatic region[J]. *Science*, 1992, **258**: 60－66.

[2] BRUNO W J, KNILL E, BALDING D J, *et al*. Effective pooling designs of library screening[J]. *Genomics*, 1995, **26**: 21－30.

[3] BALDING D J, TORNEY D C. The design of pooling experiments for screening a clone map[J]. *Fungal Genet Bio*, 1997, **21**: 302－307.

[4] PEVZNER P A. *Computational Molecular Biology: An Algorithmic Approach*[M]. Cambridge: MIT Press, 2000.

[5] HUANG T, WENG C W. Pooling spaces and non-adaptive pooling designs[J]. *Discrete Math*, 2004, **282**: 163－169.

[6] MACULA A J. Error-correcting nonadaptive group testing with d-disjunct matrices[J]. *Discrete Appl Math*, 1997, **80**: 217－232.

[7] YACHKOV A D, HWANG F K, MACULA A, *et al*. A construction of pooling designs with some happy surprises[J]. *J Comput Biol*, 2005, **12**(8): 1 127－1 134.

[8] MACULA A J. A simple construction of d-disjunct matrices with certain constant weights[J]. *Discrete Math*, 1996, **162**: 311－312.

[9] NGO H Q, DU D Z. New construction of non-adaptive and error-tolerance pooling designs[J]. *Discrete Math*, 2002, **243**: 161－170.

[10] MACULA A J. Probabilistic nonadaptive group testing in the presence of errors and DNA library screening[J]. *Ann Comb*, 1999, **3**: 61－69.

[11] 管宇, 刘越英. 单假币辨识的非适应算法[J]. 河南师范大学学报: 自然科学版, 2005, **33**(3): 19－22.

[12] NGO H Q, DU D Z. A survey on combinatorial group testing algorithms with application to DNA library screening[J]. *DMACS Series Discrete Math Theor comp sci*, 2000, **55**: 171－182.

[13] MACULA A J. Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening[J]. *J Comb Opt*, 1999, **2**: 385－397.

Pooling designs based on clone library screenings

GUAN Yu

(School of Science, Zhejiang Forestry College, Lin'an 311300, Zhejiang, China)

Abstract: A pooling design based on clone library screenings is an experimental strategy to find clones with special nucleotide strings; it is also an algorithm of combinatorial group testing. A binary conjunctional matrix, sometimes called a disjunct matrix, with a row corresponding to an experiment and a column to a clone, is usually used to represent the pooling design. Since errors exist during biological experiments, a disjunct matrix should have detecting and correcting capabilities. Utilizing the inclusion relationship between two subsets or subspaces, disjunct matrices with the same number of elements 1 in each row and column were constructed. However, the experimental number of elements was greater than the lower bound of information. [Ch, 13 ref.]

Key words: clone library; pooling design; group testing; disjunct matrix; bound of information