Probability in the Engineering and Informational Sciences, **14**, 2000, 49–56. Printed in the U.S.A.

RANDOM *k*-SET POOL DESIGNS WITH DISTINCT COLUMNS

F. K. HWANG
Department of Applied Mathematics
National Chiao-Tung University
Hsinchu, Taiwan, Republic of China
E-mail: fhwang@math.nctu.edu.tw

Random *k*-set pool designs are known to be efficient but suffer the presence of clones having identical test sets (hence indistinguishable). We propose a random pool design with distinct *k*-sets. We give an algorithm to generate the design and compute the expected numbers of unresolved negatives and unresolved positives. We show a relation between the two numbers which holds also for other random designs; hence, it can be used to obtain their expected numbers of unresolved positives which are not known. Among these other random designs, the random size-*k* design is a new design.

1. INTRODUCTION

A pool design can be represented by a $t \times n$ 0-1 matrix where each column is labeled by a clone and each row by a *pool*. A 1 entry in cell (i, j) implies that clone j is contained in pool i. For a given probe associated with a prespecified DNA fragment F, a clone is called *positive* if it contains F, and *negative* if not. The goal of a pool design is to identify the positive clones by probing the t pools. The outcome of a probing is also binary: A *negative outcome* implies that no clone in the pool is positive and a *positive outcome* implies the existence of a positive clone in the pool, but the number of such clones and their identities are not revealed.

Pool designs are also called *nonadaptive group testing algorithms* [4] which have broader applications (a probe is called a *group* and a probing is a *test* there). Typically, the literature on nonadaptive group testing algorithms assumes that tests are error-free and insists that an algorithm separates the positive items (clones) from the negative items without error. In pool designs, probing errors are well recognized

© 2000 Cambridge University Press 0269-9648/00 $12.50

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press

but ignored in most literature, as well as in this paper for easier analyses. However, ambiguities on the states of clones are allowed in a pool design. At the end of the probings, clones are classified into four categories for analysis purposes: *resolved positives*, *unresolved positives*, *resolved negatives*, and *unresolved negatives*; "unresolved" means unidentified (note that the experimenter cannot distinguish the unresolved positives from the unresolved negatives). Let P, $\bar{P}$, N, and $\bar{N}$ denote their respective numbers. For given t and n, the goal is to minimize $E(\bar{P})$ and $E(\bar{N})$, where E is the expectation function. Barillot et al. [2] first proposed $E(\bar{N})$ as the goal.

Balding et al. [1] gave an excellent survey on pool designs. There are two general categories of pool designs: *deterministic* and *random*; the latter category has the advantage of being applicable to any t and n. Among the random designs, they especially praised the effectiveness of the *random k-set* designs, first proposed by Bruno et al. [3], in which the columns are the k-subsets randomly chosen from the $\binom{t}{k}$ k-subsets of the set $\{1, \ldots, t\}$. They backed up the effectiveness claim by some examples.

Although we heartily agree that the random k-set design is an excellent choice, it is worthwhile to point out that it does have the following undesirable consequences:

i. Two columns may be identical. Then, when one of them is positive and the other negative, the design will not be able to distinguish which is which.
ii. Two rows may be identical. Then, one probing is wasted.
iii. A row can be empty (all zero). Again, a probing is wasted.

Note that the term "column" (row) is used in two ways. When we say two "columns" are identical, we refer to the column vectors in the matrix; however, when we say a "column" is positive, we refer to the clone labeling the column.

There are sound reasons to tolerate these dire consequences. In a random k-set design, the columns can be independently constructed, whereas the construction of a random k-set has routine methods. By forbidding two columns to be identical, or two rows to be identical, or a row to be empty, we destroy the independence of the columns, which will not only complicate the construction but also harden the analysis.

In this paper, we offer a partial solution by proposing the random distinct k-set pool designs in which the columns are distinct. We give an algorithm to construct such a design for any t and n. We also give an analysis along the same line given in Balding et al. [1]. Finally, we propose a relation between $E(\bar{P})$ and $E(\bar{N})$, which holds for general random designs. Although we are not doing anything to remedy consequences (ii) and (iii) *per se*, it is believed that the column distinctness will help reduce the likelihood that they will occur.

2. CONSTRUCTION OF THE RANDOM DISTINCT *k*-SET DESIGN

The naive way to forbid identical columns is to draw the k-subsets randomly, but replace a recurrent one with another k-subset. The problem of this approach is that it is time-consuming. Suppose j distinct columns have been drawn. The probability

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press

that the $(j+1)$st column randomly drawn is also distinct is

$$\left[\binom{t}{k} - j\right] \Big/ \binom{t}{k}.$$

Thus, the expected number of drawings is

$$\sum_{j=0}^{n-1} \frac{\binom{t}{k}}{\binom{t}{k} - j} = O\left(n\binom{t}{k}\right).$$

Let $f(k,t)$ denote the time required to draw a random k-subset from a t-set. Then the expected time for the naive algorithm is $O\left(n\binom{t}{k} f(k,t)\right)$, versus $O(nf(k,t))$ for the nonrestrictive design. We now give an $O\left(f\left(n,\binom{t}{k}\right)\right)$ time construction.

We first give a one–one mapping between the set of $\binom{t}{k}$ k-subsets and the set $T = \left\{1,\ldots,\binom{t}{k}\right\}$. Let $j \in T$. We transform j to a k-set $\{j_1,\ldots,j_k\}$.

Step 1. Set $x = t-1$, $y = k$, and $z = 1\left(\binom{x}{y} = 0 \text{ if } x < y\right)$.

Step 2. If $j > \binom{x}{y}$, set $j_z = t - x$, $x = x-1$, $y = y-1$, $z = z+1$, and $j = j - \binom{x}{y}$.

Step 3. If $j \leq \binom{x}{y}$, set $x = x-1$.

Step 4. Go back to Step 2. If $z = k+1$, stop.

Example: Suppose $t = 4$ and $k = 2$. The following table gives the transformations of all $j = 1,\ldots,6$. A row labeled by $>$ shows when Step 2 applies. j' is the new value of j after Step 2.

j	1	2	3	4	5	6
$>$	$\binom{1}{2}$	$\binom{2}{2}$	$\binom{2}{2}$	$\binom{3}{2}$	$\binom{3}{2}$	$\binom{3}{2}$
j_1	3	2	2	1	1	1
j'	1	1	2	1	2	3
$>$	$\binom{0}{1}$	$\binom{0}{1}$	$\binom{1}{1}$	$\binom{0}{1}$	$\binom{1}{1}$	$\binom{2}{1}$
j_2	4	4	3	4	3	2

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press

THEOREM 1: *The above algorithm yields a one–one transformation between the set of k-subsets and the set* $\left\{1,\ldots,\binom{t}{k}\right\}$.

PROOF: There are exactly $\binom{t}{k} - \binom{t-1}{k}$ k-subsets containing the index 1. The algorithm maps these k-subsets to the $\binom{t}{k} - \binom{t-1}{k}$ largest numbers in T. These k-subsets are then reduced to $(k-1)$-subsets of the set $\{2,\ldots,t\}$. The same argument applies.

With the one–one mapping available, all it takes is to draw a random n-subset from the set $\left\{1,\ldots,\binom{t}{k}\right\}$. ■

3. ANALYSIS

We assume that the number d of positive clones is given throughout this section. Since, in practice, the knowledge of d is never certain, we do not allow an automatic identification of a clone as negative solely based on the fact that d positive clones have been identified elsewhere. Our analysis essentially follows the approach of Balding et al. [1].

Row i is said to intersect column j if cell (i,j) contains a 1-entry. Call a pool with a negative outcome a *negative pool* and call the set of negative pools a *negative set*. Let $K(i)$ denote the probability that a given set of i pools is the negative set.

LEMMA 2:

$$K(i) = \sum_{h=i}^{t} (-1)^{h-i} \binom{t-i}{h-i} \left(\binom{\binom{t-h}{k}}{d} \right) \Bigg/ \left(\binom{\binom{t}{k}}{d} \right).$$

PROOF: Let H be a set of h pools containing the i given pools. There are $\binom{t-h}{k}$ distinct k-subsets not intersecting H. Hence, there are

$$\left(\binom{\binom{t-h}{k}}{d} \right)$$

ways of choosing d such k-subsets. If these d k-subsets are the positive clones, then no positive clone is contained in H. By noting that $\binom{t-i}{h-i}$ is the number of ways of choosing H, Lemma 2 follows immediately from the inclusion–exclusion principle. ■

THEOREM 3:

$$P(\bar{N} = j) = \sum_{i=0}^{t} \binom{t}{i} K(i) \left(\binom{\binom{t-i}{k}}{j} - d \right) \left(\binom{\binom{t}{k} - \binom{t-i}{k}}{n-d-j} \right) \Bigg/ \left(\binom{\binom{t}{k} - d}{n-d} \right).$$

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press

PROOF: Suppose the negative set is of cardinality i. A negative clone is unresolved if and only if none of its k indices is in the negative set. Therefore, its k indices must be chosen from the $t - i$ rows not in the negative set; but so are the k indices of a positive clone. Therefore, the number of choices for the j unresolved negatives is

$$\left(\binom{t-i}{k} - d \atop j\right).$$

On the other hand, the k indices of a resolved negative cannot all be outside of the negative set. Therefore, the number of choices for the $n - d - j$ resolved negatives is

$$\left(\binom{t}{k} - \binom{t-i}{k} \atop n-d-j\right).$$

Theorem 3 follows by averaging over i. ■

We could use $P(\bar{N} = j)$ to compute $E(\bar{N})$. However, there is a more direct argument.

THEOREM 4:

$$E(\bar{N}) = (n-d)\sum_{i=0}^{t}\binom{t}{i}K(i)\left[\binom{t-i}{k} - d\right]\Big/\left[\binom{t}{k} - d\right].$$

PROOF: Since expectation is a linear function. $E(\bar{N})$ can be obtained by multiplying $(n - d)$ to the expected value that a single negative clone is unresolved, which can be obtained by setting $n - d = j = 1$ in Theorem 3. ■

$E(\bar{P})$ was not discussed for either the random k-set design or the random incidence design (relaxing the k-set restriction). Here, we give a method to approximately compute $E(\bar{P})$ for the random distinct k-set design, which also works for the other two random designs.

We give an approximate relation between $E(\bar{P})$ and $E(\bar{N})$. To show such a relation, we need to display the parameter d as a subscript of the E function.

THEOREM 5:

$$E_d(\bar{P}) \approx d\sum_{j=0}^{n-d} P(\bar{N} = j)E_{d-1+j}(\bar{N})/[n - (d - 1 + j)].$$

PROOF: Given that $\bar{N} = j$, then a positive clone C is unresolved if and only if there does not exist a pool which contains C but no other positive clones or unresolved negatives. This is the same condition for a negative clone to be unresolved except only the nonexistence of positive clones is required there. Therefore, the former probability can be obtained from the latter by replacing d, the number of positives,

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press

with $d-1+j$, the number of other positives and unresolved negatives. Since the relation is on a per clone basis, we have to divide $E_{d-1+j}(\bar{N})$ by $n-(d-1+j)$ and to multiply by d to obtain $E_d(\bar{P})$. The reason that the relation is not exact is because the distribution of indices of C is not independent of unresolved negatives. ■

We may also compute the total number of unresolved clones from $E(\bar{N}+\bar{P}) = E(\bar{N}) + E(\bar{P})$. Explicit expressions for optimal choices of k to minimize $E(\bar{N})$, $E(\bar{P})$, and $E(\bar{N}+\bar{P})$ are still unknown.

4. RESULTS ON SOME OTHER MODELS

Theorem 5 and its proof also hold for some other random designs.

4.1. Random Incidence Design

Each clone has probability p of being included in a pool. This design was discussed in [1] and $E(\bar{N})$ given. Here, we give $P(\bar{N}=j)$, which is needed in computing $E(\bar{P})$.

THEOREM 6: *For the random incidence design*,

$$P(\bar{N}=j)=\sum_{i=0}^{t}\binom{t}{i}(1-p)^{id}[1-(1-p)^d]^{t-i}\binom{n-d}{j}$$
$$\times(1-p)^{ij}[1-(1-p)^i]^{n-d-j}.$$

PROOF: The probability that a given pool does not contain a positive clone is $(1-p)^d$. Hence, the probability that none of i given pools contains a positive clone is $(1-p)^{id}$. The probability that the other $t-i$ pools each containing a positive clone is $[1-(1-p)^d]^{t-i}$. Finally, the probability that exactly j negative clones are contained in the $t-i$ pools follows the binomial distribution as given at the end of Lemma 6. ■

The special case $j=0$ was given in [1], which also showed that $E(\bar{N})$ is minimized by setting $p=(d+1)^{-1}$.

4.2. Random *k*-Set Design

Again, $E(\bar{N})$ and $P(\bar{N}=0)$ were given in [1], but not $P(\bar{N}=j)$. Let $W(i)$ denote the probability that a given set I of i clones is the negative set.

LEMMA 7:

$$W(i)=\sum_{h=i}^{t}(-1)^{h-i}\binom{t-i}{h-i}\left[\binom{t-h}{w}\Big/\binom{t}{w}\right]^d.$$

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press

PROOF: The probability that h pools including I not containing any positive clone is

$$\binom{t-i}{h-i}\left[\binom{t-h}{w}\Big/\binom{t}{w}\right]^d,$$

where $\binom{t-i}{h-i}$ is the number of ways of choosing the other $h-i$ pools and $\binom{t-h}{w}\Big/\binom{t}{w}$is the probability that the h pools do not contain a specific positive clone. Lemma 7 follows from the inclusion–exclusion principle. ■

4.3. Random k-Size Design

This design is dual to the random k-set design in the sense that each row is a random k-subset of the set $\{1,\ldots,n\}$.

LEMMA 8:

$$P(\bar{N}=n-d)=\left[1-\binom{n-d}{k}\Big/\binom{n}{k}\right]^t,$$

$$P(\bar{N}=j)=\sum_{i=1}^{t}\binom{t}{i}\left[\binom{n-d}{k}\Big/\binom{n}{k}\right]^i\left[1-\binom{n-d}{k}\Big/\binom{n}{k}\right]^{t-i}$$
$$\times\sum_{l=j}^{n-d}(-1)^{l-j}\binom{n-d}{l}\left[\binom{n-d-l}{k}\Big/\binom{n-d}{k}\right]^i$$

for $0\le j<n-d$.

PROOF: Suppose the negative set has cardinality i. If $i=0$, then every pool contains a positive clone and no negative clone is identified. Thus, $\bar{N}=n-d$. This implies that $j<n-d$ implies $i\ge 1$. In $P(\bar{N}=j)$, the term between the two sums is the probability that the negative set is of cardinality i, whereas the term after the second sum is the probability that exactly j negative clones are not contained in any of the negative pools, computed by the inclusion–exclusion principle. ■

Note that if $i\ge 1$, then at least the k clones in a negative pool are identified; thus, $j\le n-d-k$. This fact is reflected in Lemma 8 since $\binom{n-d-l}{k}=0$ for $l>n-d-k$.

THEOREM 9:

$$E_k(\bar{N})=(n-d)\left[1-\binom{n-d-1}{k-1}\Big/\binom{n}{k}\right]^t.$$

PROOF: Given a negative clone C, the number of ways choosing a pool T containing C but not any of the d positive clones is $\binom{n-d-1}{k-1}$; namely, besides C, the other $k-1$ members of T must be chosen from the $n-d-1$ other negative clones.

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press

Therefore, the probability that there does not exist a pool containing C but no positive clones is given by the second factor of $E_k(\bar{N})$ in Theorem 9.

The k which minimizes $E_k(\bar{N})$ does not have an exact expression. If we choose k to be $n/(d+1)$ (assuming divisibility), then k equals the expected size of a random pool with optimal choice of p.

Acknowledgment

This research is partially supported by NSC grant 88-2115-M009-016.

References

1. Balding, D.J., Bruno, W.J., Knill, E., & Torney, D.C. (1996). A comparative survey of non-adaptive pooling designs. In T.P. Speed & M.S. Waterman (eds.), *Genetic mapping and DNA sequencing*. New York: Spring-Verlag, pp. 133–154.
2. Barillot, E., Lacroix, B., & Cohen, D. (1991). Theoretical analysis of library screening using a *n*-dimensional pooling strategy. *Nucleic Acids Research* 19: 6241–6247.
3. Bruno, W.J., Knill, E., Balding, D.J., Bruce, D.C., Doggett, N.A., Sawhill, W.W., Stallings, R.L., Whittaker, C.C., & Torney, D.C. (1995). Efficient pooling designs for library screening. *Genomics* 26: 21–30.
4. Du, D.Z. & Hwang, F.K. (1993). *Group testing and its applications.* Singapore: World Scientific.

https://doi.org/10.1017/S0269964800141051 Published online by Cambridge University Press