Discrete Optimization 5 (2008) 700–704

Contents lists available at ScienceDirect

Discrete Optimization

journal homepage: www.elsevier.com/locate/disopt

A new construction of $\bar{3}$-separable matrices via an improved decoding of Macula's construction

Hung-Lin Fu *, F.K. Hwang

Department of Applied Mathematics, National Chiao Tung University, Hsin Chu,Taiwan, ROC

ARTICLE INFO

Article history:
Received 30 July 2004
Received in revised form 23 March 2006
Accepted 16 April 2008
Available online 4 June 2008

Keywords:
3-separable matrices
Macula's construction

ABSTRACT

Macula proposed a novel construction of pooling designs which can effectively identify positive clones and also proposed a decoding method. However, the probability of an unresolved positive clone is hard to analyze. In this paper we propose an improved decoding method and show that for $d = 3$ an exact probability analysis is possible. Further, we derive necessary and sufficient conditions for a positive clone to be unresolved and gave a modified construction which avoids this necessary condition, thus resulting in a $\bar{3}$-separable matrix.

1. Introduction

A pooling design has many biological applications. For convenience, we use the language of clone-library screening. We have a set of n clones and a probe X which is a short DNA sequence. Let $\overline{X}$ denote the dual sequence of X, i.e., $\overline{X}$ is obtained by first reversing the order of the letters and then interchanging A with T and C with G. A clone is called *positive* if it contains $\overline{X}$ as a subsequence and *negative* if not. Typically, there are a small number of positive clones, say, from 3 to 10, among the n clones. The goal of a pooling design is to identify all positive clones through a small set of tests (or pools) performed parallelly. A test can be applied to an arbitrary subset of clones with two possible outcomes: a negative outcome indicates that the subset contains no positive clone, and a positive outcome indicates otherwise.

Let M denote the incidence matrix of a design with clones labelling the columns and tests labelling the rows. We will treat a column as the subset of row labels where i is in the column subset if and only if that column (clone) is in test i. M is called *d-disjunct* if no column is contained in the union of any other d columns. M is called *d-separable* if no two unions of distinct sets of d columns are identical, and *$\bar{d}$-separable* if d is changed to at most d. It is well known [1] that d-separable matrices can identify all positive clones if their number is exactly d, while $\bar{d}$-separable or d-disjunct matrices can identify all positive clones if their number is at most d. These matrices have become the major tools in constructing pooling designs.

Many methods have been proposed to construct these matrices. But their existence is still rare for practical need. Recently, Macula [4] opened a new door by proposing the "containment" construction method. More specifically, let $[m] = \{1, \ldots, m\}$. Then the columns of $M(m, k, d)$, $d < k$ are labelled by n random but distinct k-subsets, rows by all d-subsets, and $M_{ij} = 1$ if and only if the row label is contained in the column label. Macula proved that M is d-disjunct. This approach extends to many other containment relations as in partial orders [2] and geometrical structures [6].

One problem with this construction is that the number n of columns is bounded by $\binom{m}{k}$, hence m cannot be too small. On the other hand, the number of tests is $\binom{m}{d}$, so d must be small. Macula [5] proposed using $M(m, k, 2)$even though the actual

* Corresponding author.
E-mail address: hlfu@math.nctu.edu.tw (H.-L. Fu).

1572-5286/$ – see front matter
doi:10.1016/j.disopt.2008.04.002

number d of positive clones can be larger than 2. In such an application there may exist "unresolved" clones whose status of being positive or negative is unknown.

Let P^+ denote the probability that a positive clone is unresolved. The problem of computing P^+ under a given decoding method turns out to be difficult. Macula [5] gave a simple decoding while Hwang and Liu [3] improved it, but with a more complicate analysis where P^+ can be computed only for $d = 3$.

In this paper we further improve the decoding method. Although probability analysis remains difficult, we are able to accomplish the following for $d = 3$:

(1) obtain necessary and sufficient conditions for a positive clone to be unresolved;
(2) give an exact probability analysis; and
(3) derive a simple necessary condition and show that by choosing the column indices judiciously, this necessary condition can be avoided and hence the matrix obtained is $\bar{3}$-separable.

2. The unique-graph decoding

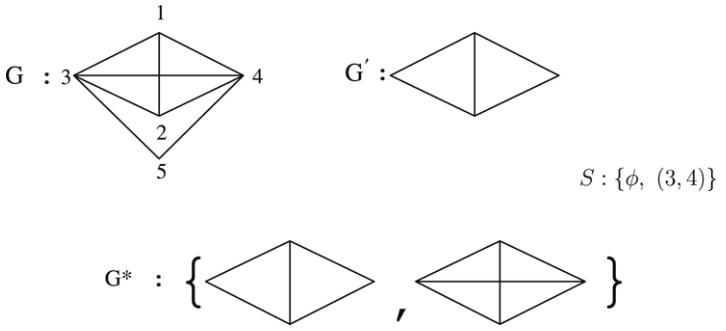
Consider $M(m, k, 2)$ throughout this section. We shall use (u, v) to denote an edge joining u and v. The *outcome graph* G has $[m]$ as its vertex-set and an edge (u, v) if the pool labelled by (u, v) is positive, i.e., it has a positive outcome. Note that each positive clone induces a k-clique in G, but a k-clique in G can correspond to a negative clone (or a k-subset of $[m]$ not chosen as a column label) while all its edges actually come from some other positive clones.

The *unique-graph decoding* consists of the following rules:

(i) If a clone appears in a negative pool, it is negative.
(ii) If a k-clique in G contains an edge not appearing in any other k-clique (in G), then it represents a positive clone.
(iii) Let p' be the number of positive clones identified in (ii). Let G' be obtained from G by removing all p' k-cliques identified in (ii) as positive clones and also removing isolated vertices. Let S denote the set of edges in these k-cliques with both endpoints in G'. Define $G^* = \{G' \cup S^* : S^* \subseteq S\}$. If G^* contains a unique graph $G' \cup S^*$ which is the union of a unique set of p'' k-cliques for some p'' with $p' + p'' \leq p$, then each of these p'' k-cliques represents a positive clone.
(iv) All clones not identified in (i), (ii), (iii) are unresolved.

Note that rule (iii) differentiates the unique-graph decoding from the original Hwang–Liu decoding. The following example illustrates the difference.

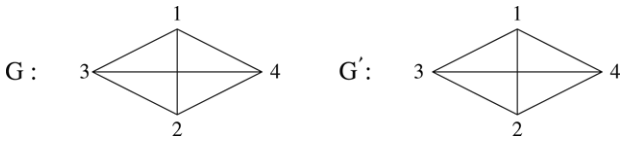
Example 1. $k = 3$, $d = 3$, $\Delta 123$, $\Delta 124$, $\Delta 345$ are positive.



Using (ii), only $\Delta 345$ is identified (since either (3,5) or (4,5) is an edge not in any other triangle). Using (iii), G^* consists of two graphs while only the first one is the union of two triangles. Thus $\Delta 123$ and $\Delta 124$ are identified.

Even the unique-graph decoding can leave positive clones unresolved.

Example 2. $k = 3$, $d = 3$.



Since G' is the union of four different sets of three triangles, the unique-graph decoding fails to identify any positive clone. Define $A \oplus B = (A \setminus B) \cup (B \setminus A)$. We have:

Theorem 2.1. *Under the unique-graph decoding, a positive clone A is unresolved if and only if there exists a nonpositive k-clique B such that all edges of $A \oplus B$ are contained in the other positive clones.*

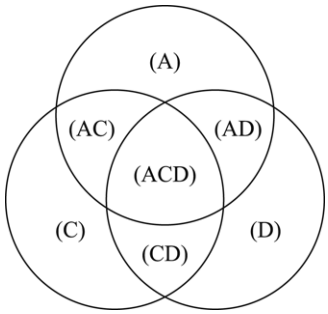


Fig. 1. Intersection of A, C, D.

Proof. "if ". We cannot tell whether A or B is positive since they induce the same outcome graph.

"only if." Suppose there does not exist a nonpositive k-clique in G. Then every k-clique in G must be positive. On the other hand, every positive k-clique is clearly in G. Therefore the number of k-cliques in G equals the true number of positive k-cliques; hence at most p. Consequently, these positive k-cliques can be identified through rule (iii) of the unique-graph decoding.

The requirement that all edges in $A \setminus B$ are contained in the other positive clones follows from the fact that if a k-clique K contains an edge which is not in any other k-clique, then K represents a positive clone. Finally, all edges in $B \setminus A$ are necessarily in the other positive clones since otherwise B would not be in G. □

Corollary 2.2. *If $d = 3$, then a necessary and sufficient condition for A to be unresolved is that each of the other two positive clones contains $A \oplus B$.*

Proof. Let C and D denote the other two positive clones. For $a \in A \setminus B$, $C \cup D$ must contain all edges from a to $A \setminus \{a\}$. But neither C nor D alone can contain all of them since it requires $C \supseteq A$ or $D \supseteq A$, an absurdity. Therefore $a \in C$ and $a \in D$. Similarly, we can prove that every $b \in B \setminus A$ is in both C and D. □

Corollary 2.3. *For $d = 3$, a necessary condition for A to be unresolved is that G contains a k'-clique with $k' > k$.*

Proof. $A \cup B \setminus A = B \cup A \setminus B$ is a k'-clique with $k' > k$. □

3. Exact probability analysis for $d = 3$

The necessary and sufficient condition in Theorem 2.1 is not convenient for computing the probabilities of unresolved clones since it involves an unspecified negative clone. For $d = 3$, we transform the condition into conditions involving the three positive clones A, C, D only. Fig. 1 shows the intersections of A, B, C where the seven parts are labelled by $(A), (C), (D), (AC), (AD), (CD)$ and (ACD).

A necessary condition for A to be unresolved is that $(A) = \phi$, which implies $(AC) \neq \phi$, $(AD) \neq \phi$ (otherwise C or D would contain A, an absurdity). Furthermore (CD) cannot be empty since otherwise the edge(s) from (AC) to (AD) forces A to be positive. Finally, by Theorem 2.1 all edges from (AC) to (AD) must be in B, which implies $B \supset (AC) \cup (AD)$. So if B contains x vertices in (CD), B must leave x vertices in (ACD) out to enforce $|A| = |B| = k$.

Let P_n^+ denote the probability a given positive clone is unresolved.

Theorem 3.1. *Consider three positive clones A, C, D. Then A is unresolved if and only if (A) is empty while $|(CD) \cup (ACD)| \geq 2$.*

Proof. "only if." Shown in the preceding paragraph.

"if." Note that $(A) = \varnothing$ and $|(CD)| \geq 1$ implies $|(AC) \cup (AD) \cup (CD) \cup (ACD)| \geq k + 1$. B can be selected from $(AC) \cup (AD) \cup (CD) \cup (ACD)$ with the provision that if B takes x vertices from (CD), then it must leave x vertices in (ACD) untouched so that $|A \setminus B| = |B \setminus A|$. □

We are now ready to give the probability formulas.

Define $m = \min\{k - i, i - j\}$.

Theorem 3.2. $P_n^+ = \sum_{i=1}^{k-1} \binom{k}{i} \binom{b-k}{k-i} \sum_{j=0}^{i-1} \binom{i}{j} \sum_{h=(2-j)^+}^{m} \binom{k-i}{h} \binom{b-2k+i}{i-j-h} / \binom{N-1}{2}$.

Proof. $\binom{k}{i} \binom{b-k}{k-i}$ is the number of choices of C with $|(AC) = i|$. The rest of the numerator gives the number of choices of D such that D takes all the $k - i$ vertices of $A \setminus C$ plus j vertices of (AC), h vertices of $C \setminus A$, and the rest outside of $A \cup C$. Note that $j + h \geq 2$. The denominator gives the unconstrained number of choices of C and D (if we also consider the choice of the $n - 3$ negative clones, we merely add the term $\binom{N-3}{n-3}$ to both the numerator and denominator). □

Note that P_n^+ is independent of n.

Let $P_n^+(x)$ denote the probability that exactly x positive clones are unidentified.

Lemma 3.3. *$x = 3$ if and only if G is a k'-clique with $k' > k$.*

Proof. Clearly, a trivial necessary condition for $x = 3$ is that $(A) = (C) = (D) = \phi$, or $G = A \cup C \cup D$ is a k'-clique, with $k' > k$. We now show that this condition is also sufficient.

Suppose $|(ACD)| = j$. Then, $|(AC)| = |(AD)| = |(CD)| = (k - j)/2$. Therefore $|(CD)| + |(ACD)| = j + \frac{k-j}{2} = \frac{k+j}{2} \geq 2$ ($k = 3$ forces $j = 1$), satisfying the condition of Theorem 3.1. Hence A is unresolved. Similarly, we can prove that C and D are unresolved. □

Corollary 3.4. $P_n^+(3) = \sum_{j=0}^{k-1} \binom{k}{j} \binom{k-j}{(k-j)/2} \binom{b-k}{(k-j)/2} \Big/ \binom{N}{3}$.

Proof. Given A, there are $\binom{k}{j}$ choices of j vertices in (ACD), and $\binom{k-j}{(k-j)/2}$ choices of $(k - j)/2$ vertices in AC and $\binom{b-k}{(k-j)/2}$ choices for the remaining vertices of C. Once A and C are chosen, D is fixed. □

Lemma 3.5. *$x = 2$ if and only if exactly two of (A), (C) and (D) are empty, say, C and D, and $|(CD)| \leq k - 2$.*

Proof.

$$|(AC)| = k - |(ACD)| - |(CD)| = |(AD)| \geq 1,$$

since otherwise $C = D$, an absurdity. Further,

$$|(AC)| + |(ACD)| = |(AD)| + |(ACD)| = k - |(CD)|.$$

Lemma 3.5 follows from Theorem 3.1 immediately. □

Corollary 3.6. $P_n^+(2) = 3 \sum_{i=2}^{\lfloor (k-1)/2 \rfloor} \binom{k}{i} \sum_{j=2i-k+1}^{i-1} \binom{i}{j} \binom{k-i}{i-j} / \binom{N-1}{2}$.

Proof. Suppose A is the only identified positive clone. Assume $|(AC)| = i$ and $|(ACD)| = j$. Then $|(AD)| = i$ as shown in the proof of Lemma 3.5 Since A is identified,

$$k = |(A)| > |(AC)| + |(AD)| + |(ACD)| = 2i - j.$$

Hence

$$j \geq 2i - k + 1,$$

and

$$2i \leq k - 1.$$

Note that we set $i \geq 2$ to guarantee $|(CD)| = k - i \leq k - 2$. The multiplication by 3 is because any of A, C, D can be the unique resolved one. □

Let $(P_n^+)'$ be obtained from P_n^+ by changing the upper bound of h from m to $m - 1$ in Theorem 3.2 to guarantee that neither (C) nor (D) is empty. Then $(P_n^+)'$ is the probability that A is the unique unresolved positive clone. Hence:

Lemma 3.7. $P_n^+(1) = 3(P_n^+)'$.

Theorem 3.8. $P_n^+(0) = 1 - P_n^+(3) - P_n^+(2) - P_n^+(1)$.

Note that $P_n^+(x)$ is independent of n for any x.

4. A new construction of $\bar{3}$-separable matrices

Theorem 4.1. *Suppose $d \leq 3$ and G contains no $(k + 1)$-clique. Then the unique-graph decoding identifies all d positive clones.*

Proof. Suppose $d = 1$ or 2. Since a single positive clone cannot cover the edges of another positive clone, rule (ii) of the unique-graph decoding always identifies all positive clones. Suppose $d = 3$. Then the proof follows from Corollary 2.3. □

We now show how to choose the labels of the columns such that G does not contain a $(k + 1)$-clique.

Partition the m indices of $[m]$ evenly into k parts. Then K is a *legitimate label* of columns if K consists of one index from each part. Now, a $(k + 1)$-clique has $k + 1$ indices, hence two of which, say, u and v, must come from the same part. But no label containing both u and v is chosen since it is not legitimate. Therefore G does not contain the edge (u, v), and *à* fortiori, does not contain the $(k + 1)$-clique containing u and v.

Theorem 4.2. *The number of legitimate labels is approximately* $(\frac{m}{k})^k$.

Therefore, by Theorem 4.1, we have a construction of a $\bar{3}$-separable matrix with $\binom{m}{2}$ tests and $(\frac{m}{k})^k$ clones. We now explain why this construction does not lead to a 3-disjunct matrix. Note that rule (iii) of the unique-graph decoding identifies positive clones even when there exist unresolved nonpositive k-cliques (see Example 1, where triangles 124 and 134 are nonpositive and unidentified). On the other hand, a d-disjunct matrix has the property that all clones which are not identified as negative are identified as positive, a clear conflict with rule (iii).

The following comparison shows the advantage and disadvantage respectively of using this new construction. Given n clones, we want to choose (m, k) such that

$$n \le \left(\frac{m}{k}\right)^k \tag{1}$$

with the minimum m (so the number of tests is minimized). Approximate (1) by equality. Then

$$m = kn^{\frac{1}{k}}. \tag{2}$$

To minimize the right-hand side size of (2) with respect to k, we obtain $k^0 = \ln n$. Consequently, $m^0 = e \ln n$. So the number of tests is $\binom{e \ln n}{2}$.

To compare with $M(m, k, 2)$, we first choose (m, k) such that

$$\binom{m}{k} \ge n \tag{3}$$

and m is minimum. Since $k = \lfloor \frac{m}{2} \rfloor$ maximizes $\binom{m}{k}$, we replace k by $\lfloor \frac{m}{2} \rfloor$ in (3). Denote by m_n the m minimizing the modified (3). There the number of tests is $\binom{m_n}{2}$.

Example. Let $n = 1000$. Then $m^0 = 20$ and $k^0 = 6$ since $4 + 4 + 3 + 3 + 3 + 3 = 20$, and $4^2 3^4 > 1000$.

Thus our $\bar{3}$-separable matrix requires $\binom{20}{2} = 190$ tests.

On the other hand, $m_{1000} = 13$ ($k = 6$) since $\binom{13}{6} > 1000 > \binom{12}{6}$.

Hence $M(13, 6, 2)$ requires 78 tests. However, to identify three positive clones, we need to use $M(13, 6, 3)$ which requires 286 tests.

Acknowledgments

The authors would like to thank the referees for their helpful comments and suggestions.

References

[1] D.Z. Du, F.K. Hwang, Combinatorial Group Testing and its Applications, 2nd ed., World Scientific, Singapore, 2000.
[2] T.Y. Huang, C.W. Weng, Pooling spaces and nonadaptive pooling designs, Discrete Math. 283 (2004) 163–169.
[3] F.K. Hwang, Y.C. Liu, Random pooling designs under various structures, J. Comb. Optim. 7 (2003) 339–352.
[4] A.J. Macula, A simple construction of d-disjunct matrices with certain constant weights, Discrete Appl. Math. 50 (1996) 217–222.
[5] A.J. Macula, Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening, J. Comb. Optim. 2 (1999) 385–397.
[6] H. Ngo, D.Z. Du, New constructions of nonadaptive and error-tolerance pooling designs, Discrete Math. 243 (2002) 161–170.