TAIWANESE JOURNAL OF MATHEMATICS
Vol. 10, No. 4, pp. 927-932, June 2006
This paper is available online at http://www.math.nthu.edu.tw/tjm/

BOUNDING THE NUMBER OF COLUMNS WHICH APPEAR ONLY IN POSITIVE POOLS

H. B. Chen, F. K. Hwang and C. M. Li

Abstract. d-separable, $\overline{d}$-separable and d-disjunct matrices are the major tools in constructing pooling designs which has many applications to DNA experiments, for example, the clone library screening problem. While there exists a simple decoding for d-disjunct matrices, only brute-force methods are known for the other two. In this paper we identify structures in these two matrices which lead to significant improvements for decoding.

1. Introduction

Nonadaptive group testing has been intensively studied recently due to its biological applications. For a set C of given objects consisting of positive and negative ones, a group test is performed on an arbitrary subset $S \subseteq C$ with two possible outcomes: a positive outcome signifies that S contains a positive object and a negative outcome signifies otherwise. The goal is to identify all positive objects through a minimum number of group tests. In biological experiments, the objects are usually clones and a clone is positive if it contains a prespecified DNA segment. To save experiment time, it is important that all test subsets are specified before any testing is done, known as *nonadaptive group testing* in the group testing literature, and *a pooling design* in biological applications (each test-subset is a pool).

Let M denote the incidence matrix of a pooling design with rows as pools and columns as clones. We will view a column as the set of locations of its 1-entries, i.e., a column is a subset of the set of row indices. We will use the terminology *a positive (negative) column* to mean that the column represents a positive (negative)

Received April 22, 2004, accepted December 13, 2004.
Communicated by Xuding Zhu.
2000 *Mathematics Subject Classification*: 05C35, 05C85.
Key words and phrases: Pooling design, Nonadaptive group testing, d-separable matrix, $\overline{d}$-separable matrix.
This research is partially supported by Republic of China, National Science Council grant NSC 92-2115-M-009-014.

object, and *a positive (negative) pool* to mean that the test has a positive (negative) outcome. Let V denote the (binary) outcome vector, i.e., $v_i = 1(0)$ if row i is positive (negative).Then V can be interpreted as the union of all positive columns.

Three types of binary matrices have become the major tools in constructing a pooling design:

(i) M is *d-separable* if no two unions of d columns are same.

(ii) M is *$\overline{d}$-separable* if no two unions of at most d columns are same.

(iii) M is *d-disjunct* if no column is contained in the union of any other d columns.

Let p denote the actual (unknown) number of positive objects. It is well known [2] that the d-separable matrix can identify all p positive clones if $p = d$, and the $\overline{d}$-separable matrix or the d-disjunct matrix can identify all p positive clones if $p \leq d$ (the d-disjunct matrix also has a simple decoding). These matrices have also been studied in extremal set theory [5, 7, 8] and coding [4, 10], other than the biological applications.

Let M denote a d-separable or $\overline{d}$-separable or d-disjunct matrix. We will bound the number of columns not appearing in any negative pool. Later, we consider the same with an additional constraint. We give two applications of our results in the decoding of positive clones in a pooling design.

2. Main Results

Let M be a $t \times n$ d-separable or $\overline{d}$-separable or d-disjunct matrix, $\{D_1, \cdots, D_p\}$ the set of positive clones, and V the outcome vector corresponding to $\{D_1, \cdots, D_p\}$, i.e., $V = \bigcup_{i=1}^{p} D_i$. Let T_0 and T_1 denote the sets of negative pools and positive pools, respectively, with $|T_0| = t_0$ and $|T_1| = t_1$, $t_0 + t_1 = t$. Let $M_1(M_0)$ be the $t_1(t_0) \times n_1(n_0)$ submatrix of M such that the rows are $T_1(T_0)$ and the columns are those which have no 1-entries in $T_0(T_1)$.

Lemma 2.1. *M is d-separable (or $\overline{d}$-separable or d-disjunct) implies M_1 is d-separable (or $\overline{d}$-separable or d-disjunct).*

Proof. Obvious from the fact that a column in M_1 preserves all the 1-entries in M, hence M_1 inherits the property of M. ∎

An immediate consequence of Lemma 2.1 is that we can use bounds of n for a d-separable(or $\overline{d}$-separable or d-disjunct) matrix to bound $n_0(n_1)$. For a d-disjunct matrix, Füredi [8] proved $n \leq d \cdot 2^{4t/d^2}$ by a combinatorial argument. D′yachkov and Rykov [4] gave the asymptotic bound $n \leq d \cdot 2^{2t/d^2}(1 + o(1))$.

Bounds of n_1 for d-separable and $\overline{d}$-separable matrices can be obtained through their relation with d-disjunct matrices. Kautz and Singleton [10] proved that a $\overline{d}$-separable matrix is a $(d-1)$-disjunct matrix. Thus

Theorem 2.2. *Suppose M is $\overline{d}$-separable. Then $n_1 \leq (d-1) \cdot 2^{4t_1/(d-1)^2}$. Also $n_1 \leq (d-1) \cdot 2^{2t_1/(d-1)^2}(1+o(1))$ asymptotically.*

Recently, Chen and Hwang [1] proved that a d-separable matrix can be converted to a $\lfloor d/2 \rfloor$-disjunct matrix by adding a row. Thus

Theorem 2.3. *Suppose M is d-separable. Then $n_1 \leq \lfloor d/2 \rfloor \cdot 2^{4(t_1+1)/\lfloor d/2 \rfloor^2}$. Also $n_1 \leq \lfloor d/2 \rfloor \cdot 2^{2(t_1+1)/\lfloor d/2 \rfloor^2}(1+o(1))$ asymptotically.*

Ironically, the M being d-disjunct case does not have any analogous result. This is because that a much stronger result is well known. Suppose the actual number of positive clones is $p \leq d$. Then $n_1 = p$ [10].

Note that M_1 actually satisfies an additional constraint that there exists a set D of d columns in M_1 such that the union of D intersects all rows in M_1(any set of d columns containing all positive clones will do). We will make use of this constraint to derive a new bound for the d-separable case.

Let N_1 denote the set of columns in M_1 and let $D = \{D_1, \cdots, D_d\}$. Define $D_i^* = D_i \setminus \bigcup_{j \neq i} D_j$ for $1 \leq i \leq d$.

Lemma 2.4. *$C \cap D_i^* \neq C' \cap D_i^*$ for all C, $C' \in N_1 \setminus D$ and $1 \leq i \leq d$.*

Proof. Suppose to the contrary that there exist C, C' and i such that

$$C \cap D_i^* = C' \cap D_i^*.$$

Since the union of D intersects all rows in M_1,

$$C \setminus D_i^* \subseteq \bigcup_{j \neq i} D_j \text{ and}$$

$$C' \setminus D_i^* \subseteq \bigcup_{j \neq i} D_j.$$

Thus we have $C \cup (\bigcup_{j \neq i} D_j) = C' \cup (\bigcup_{j \neq i} D_j)$, violating the assumption of d-separability. ■

Theorem 2.5. *$n_1 \leq d + 2^{\lfloor t_1/d \rfloor} - 1$ for M d-separable.*

Proof. Clearly,

$$\min_{1\le j\le d} \mid D_j^* \mid \le \lfloor t_1/d \rfloor.$$

Without loss of generality, assume D_i^* achieves the minimum. By Lemma 2.4, all columns in $N_1 \setminus D$ have distinct intersections with D_i^*, hence there are at most $2^{|D_i^*|} \le 2^{\lfloor t_1/d \rfloor}$ of them. But we have to subtract one since the intersection number cannot be $\mid D_i^* \mid$, or the union of that column with $\bigcup_{j\neq i} D_j$ equals D, violating the assumption of d-separability. ∎

Compare the two bounds in Theorems 2.3 and 3.5, the bound in Theorem 2.5 is better for $d \le 16$, which is usually the case in biological applications.

Theorem 2.6. *For a $\overline{d}$-separable matrix M,*

$$n_1 \begin{cases} = p & \text{if } p \le d-1, \\ \le d + 2^{\lfloor t_1/d \rfloor} - 2 & \text{if } p = d. \end{cases}$$

Proof. Since M_1 is $\overline{d}$-separable, hence $(d-1)$-disjunct, the only columns in M_1 are the positive clones if $p \le d-1$.

If $p = d$, then $C \cap D_i^*$ can be neither D_i^* nor $\emptyset$ (leading to $C \cup (\bigcup_{j\neq i} D_j) = \bigcup_{j\neq i} D_j$). Hence 2 is subtracted from $2^{\lfloor t_1/d \rfloor}$. ∎

3. Two Applications

As mentioned before, the d-disjunct matrix comes with a simple decoding, namely, a column is positive if and only if it does not appear in a negative row. On the other hand, the d-separable matrix and the $\overline{d}$-separable matrix have fewer tests but have no simple decoding. The only known decoding is the brute-force method [11] by computing the output vectors of all candidate sets of positive clones (this can be done in advance) and check which matches the actual outcome vector. Let S and $\overline{S}$ denote the sizes of the candidate sets for d-separable and $\overline{d}$-separable, respectively. Then essentially,

$$S = \binom{n_1}{d} \text{ and } \overline{S} = \sum_{j=0}^{d} \binom{n_1}{j}.$$

Our results show that n can be replaced by the bounds of n' in Theorem 2.3 or 2.5 in S, and by the bounds in Theorem 2.2 or Theorem 2.6 in $\overline{S}$ for large savings.

In some biological applications, besides the positive and negative objects, there is a third category called *inhibitors* where the presence of an inhibitor in a pool

dictates its outcome to be negative regardless of how many positive clones are present. Such a model was first proposed in [6].

Assume there are at most d positive clones and at most r inhibitors, it was shown [3, 9] that a $(d+r)$-disjunct matrix can identify all positive clones. However, the decoding for positive clones requires inspecting all $\binom{n'}{r}$ r-subsets of the n' columns which are candidates of inhibitors. In [3], n' was just set to n. In [9], n' was somewhat reduced, but no upper bound was derived.

Note that an inhibitor cannot appear in a positive pool. So a column is a candidate of inhibitor if it does not intersect T_1. By Lemma 2.1, M_0 is $(d+r)$-disjunct, hence $n' \leq (d+r) \cdot 2^{4t/(d+r)^2}$ or $n' \leq (d+r) \cdot 2^{2t/(d+r)^2}(1+o(1))$. The reason that the bound in Theorem 2.5 is not applicable is that the set of inhibitors does not necessarily span T_0, which is needed in the proof of Lemma 2.4. Namely, T_0 can contain a row consisting of negative clones but no inhibitor.

REFERENCES

1. H. B. Chen and F. K. Hwang, Exploring the missing link among d-separable, $\overline{d}$-separable and d-disjunct matrices, preprint, 2003.
2. D. Z. Du and F. K. Hwang, Combinatorial Group Testing and Its Applications, 2^{nd} ed., World Scientific, Singapore, 2000.
3. A. G. D′yachkov, A. J. Macula, D. C. Torney and P. A. Villenkin, Two models of nonadaptive group testing for decoding experiments, in A. C. Attkinson, P. Hackl and W. G. Muller (eds): Proc. 6^{th} Inter. Workshop in Model Oriented Design and Analysis, *Physica-Verlog*, 2001, pp. 63-75.
4. A. G. D′yachkov and V. V. Rykov, Bounds of the length of disjunct codes, *Probl. Control Inform. Thy.*, **11** (1982), 7-13.
5. P. Erdös, P. Frankl and Z. Füredi, Families of finite sets in which no set is covered by the union of r others, *Israel J. Math.*, **51** (1985), 79-89.
6. M. Farach, S. Kannan, E. Knill and S. Muthukrishnan, Group testing problems with sequence in experimental molecular biology, Proc. Compresson and Complexity of Sequences, B. carpentieri et al. (eds), IEEE Press, pp. 357-367, 1997.
7. P. Frankl and Z. Füredi, Union-free Hypergraphs and probability Theory, *Euro. J. Combin.*, **5** (1984), 127-131.
8. Z. Füredi, On r-cover-free families, *J. Combin. Thy.*, A **75** (1996), 172-173.
9. F. K. Hwang and Y. C. Liu, Error-tolerant pooling designs with inhibitors, *J. Comput. Biology*, **10** (2003), 231-236.
10. W. H. Kautz and R. R. Singleton, Nonrandom binary superimposed codes, *IEEE Trans. Inform. Thy.*, **10** (1964), 363-377.

11. W. Wu, C. Li, X. Wu and X. Huang, Decoding in pooling designs, *J. Combin. Opt.*, **7** (2003), 385-388.

H. B. Chen, F. K. Hwang and C. M. Li
Department of Applied Mathematics,
National Chiao Tung University,
Hsinchu 300, Taiwan, R.O.C.