Bull. Aust. Math. Soc. **85** (2012), 121–127
doi:10.1017/S0004972711002565

A NEW CONSTRUCTION FOR POOLING DESIGNS

FENGLIANG JIN✉, HOUCHUN ZHOU and JUAN XU

(Received 18 May 2011)

Abstract

Pooling designs are a very helpful tool for reducing the number of tests for DNA library screening. A disjunct matrix is usually used to represent the pooling design. In this paper, we construct a new family of disjunct matrices and prove that it has a good row to column ratio and error-tolerant property.

2010 *Mathematics subject classification*: primary 05C10.

Keywords and phrases: disjunct matrix, pooling design, row to column ratio, error-tolerant property.

1. Introduction

The basic problem of group testing is to identify the set of defective items in a large population of items. Group testing algorithms can roughly be divided into two categories: combinatorial group testing (CGT) and probabilistic group testing (PGT). In CGT, it is often assumed that the number of positives among n items is equal to or at most d for some given positive integer d. In PGT, we fix some probability p of having a positive. Group testing strategies can also be either *adaptive* or *nonadaptive*. A group testing algorithm is nonadaptive if all tests must be specified without knowing the outcomes of other tests. A nonadaptive testing algorithm is useful in many areas such as DNA library screening. 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 group testing algorithm is *error tolerant* if it can detect some errors in test outcomes.

A binary incidence matrix, sometimes called a disjunct matrix, with a row corresponding to an experiment and a column corresponding to a clone, is usually used to represent the pooling design. Kautz and Singleton [6] were first to propose the concept of a d-disjunct matrix. Macula [7] proposed a novel way of constructing d-disjunct matrices based on the containment relation of subsets in a finite set. As a generalization of Macula's construction, Zhao [10] constructed a family of disjunct matrices and discussed its error-tolerant property.

This work is partially supported by grants from the National Natural Science Foundation of China (NSFC No. 10771120) and the Shandong Natural Science Foundation of China (No. Y2008A27).

However, when there are errors in the test outcomes, the design no longer works. To deal with this case, Macula [8] proposed a d^e-disjunct matrix which is a mathematical model of error-tolerance design. D'yachkov *et al.* [2] proved that a d^e-disjunct matrix can detect $e-1$ errors and correct $\lfloor (e-1)/2 \rfloor$ errors. D'yachkov *et al.* [3] discussed the error-tolerant property of Macula's construction. Ngo and Du [9] proposed a family of d-disjunct matrices based on matchings of the complete graph K_{2m}. Bai *et al.* [1] generalized Ngo and Du's construction, and obtained two families of d^e-disjunct matrices based on the substructures of Johnson graphs and Grassmann graphs. Huang and Weng [5] generalized Ngo and Du's constructions to pooling spaces, and proved that a d^{2e}-disjunct matrix is e-error-correcting in [4].

The rest of this paper is organized as follows. Section 2 presents basic notations and related works. Section 3 proposes a new construction of disjunct matrix based on an n-partite complete graph $G_{m,m,\dots,m}$ and discusses its row to column ratio and error-tolerant property.

2. Preliminaries

In this paper, for any positive integer v we shall use $[v]$ to denote $\{1, 2, \dots, v\}$. Also, given any set X and integer k, $\binom{X}{k}$ denotes the collection of all k-subsets of X.

For a 0–1 matrix M, a row corresponds to a test (pool) and a column corresponds to a clone. If $M_{ij} = 1$ then clone j is contained in pool i. The weight of a row or a column is the number of 1s it has. For $t+1$ distinct columns of M, namely C_0, $C_1, \dots, C_t$, if $C_0 \le C_1 + \cdots + C_t$ (the '+' represents Boolean summation: $0+0=0$, $0+1=1+0=1+1=1$), it is said that C_0 is covered by $C_1, \dots, C_t$.

Definition 2.1 [6]. We say M is d-disjunct if the union of any d columns does not contain another column.

Lemma 2.2 [9]. *The matrix M is d-disjunct if and only if for any set of $d+1$ distinct columns $C_{j0}, C_{j1}, \dots, C_{jd}$ with one column (say, C_{j0}) designated, C_{j0} has a* 1 *in some row where all C_{jk}, $1 \le k \le d$, contain* 0*s.*

Let $S(\overline{d}, n)$ denote the set of all subsets of n items (or columns) with size at most d, called the set of *samples*. For $s \in S(\overline{d}, n)$, let $P(s)$ denote the union of all columns corresponding to s, that is, $P(s) = \bigcup_{i \in s} C_i$. A pooling design is e-error-detecting (correcting) if it can detect (correct) up to e errors in test outcomes. In other words, if a design is e-error-detecting then the test outcome vectors form a v-dimensional binary code with minimum Hamming distance at least $e+1$. Similarly, if a design is e-error-correcting then the test outcome vectors form a v-dimensional binary code with minimum Hamming distance at least $2e+1$. The following remarks are simple to see, and will be useful later on.

Remark 2.3 [9]. Suppose that M has the property that for any $s, s' \in S(\overline{d}, n)$, $s \ne s'$, $P(s)$ and $P(s')$ viewed as vectors have Hamming distance k or greater. In other words, $|P(s) \oplus P(s')| \ge k$ where $\oplus$ denotes the symmetric difference. Then M is $(k-1)$-error-detecting and $\lfloor (k-1)/2 \rfloor$-error-correcting.

https://doi.org/10.1017/S0004972711002565 Published online by Cambridge University Press

DEFINITION 2.4 [8]. We say M is d^e-disjunct if given any $d+1$ distinct columns with one designated, there are $e+1$ rows with a 1 in the designated column and 0 in each of the other d columns.

Obviously, a d^e-disjunct matrix with $e=0$ is said to be d-disjunct. For a d^e-disjunct matrix, the smaller the row to column ratio, the better the design; and the larger e is, the better the design is. So the basic problem of pooling designs is to construct a disjunct matrix such that its row to column ratio is small and e is large.

In the following, we give some related work about constructions of disjunct matrices over graphs.

Macula [7] proposed a novel way of constructing a family of d-disjunct matrices of order $\binom{n}{d} \times \binom{n}{k}$ with row weight $\binom{n-d}{k-d}$ and column weight $\binom{k}{d}$.

DEFINITION 2.5 [7]. For positive integers $1 \le d < k < n$, let $M(d,k,n)$ be the binary matrix with row (respectively, column) indexed by $\binom{[n]}{d}$ (respectively, $\binom{[n]}{k}$) such that $M(A,B) = 1$ if and only if $A \subseteq B$ and 0 otherwise.

Ngo and Du [9] constructed a $g(m,d) \times g(m,k)$ d-disjunct matrix $M(m,k,d)$ with row weight $g(m-d, k-d)$ and column weight $\binom{k}{d}$, where $g(m,l) = \binom{2m}{2l}(2l)!/2^l l!$. Furthermore, $M(m,m,d)$ is d^d-disjunct and can detect d errors and correct $\lfloor d/2 \rfloor$ errors. A matching of size l (that is, it has l edges) is called an l-matching and the matrix of Ngo and Du is constructed as follows.

DEFINITION 2.6 [9]. For positive integers $1 \le d < k \le m$, let $M(m,k,d)$ be the 0–1 matrix whose rows are indexed by the set of all d-matchings on K_{2m}, and whose columns are indexed by the set of all k-matchings on K_{2m}. All matchings are to be ordered lexicographically. Then $M(m,k,d)$ has a 1 in row i and column j if and only if the ith d-matching is contained in the jth k-matching.

Zhao [10] generalized Macula's construction and constructed a $\binom{n}{d}m^d \times \binom{n}{k}m^k$ d-disjunct matrix with row weight $\binom{n-d}{k-d}m^{k-d}$ and column weight $\binom{k}{d}$. Let G denote the n-partite complete graph $G_{m,m,\ldots,m}$ and G_k denote the set of all complete subgraphs of G on k vertices.

DEFINITION 2.7 [10]. For positive integers $1 \le d < k < n$, let $M(d,k,n;m)$ be the binary matrix with row (respectively, column) indexed by G_d (respectively, G_k) such that $M(D,K) = 1$ if and only if $D \subseteq K$ and 0 otherwise.

3. Main results

The research summarized in the previous section stimulated us to construct a new family of disjunct matrices based on the complete subgraphs of a multipartite complete graph.

Let G denote the n-partite complete graph $G_{m,m,\ldots,m}$ and K_n denote a complete subgraph of G on n vertices. Recall that two graphs are disjoint if they have no vertices in common. Let H_l denote a set of l pairwise disjoint complete subgraphs of G on n vertices.

https://doi.org/10.1017/S0004972711002565 Published online by Cambridge University Press

Definition 3.1. For positive integers $1 \le d < k \le m$, let $M(d, k, m; n)$ be the binary matrix whose rows (respectively, columns) are indexed by the set of all H_d (respectively, H_k). Then $M(d, k, m; n)$ has a 1 in row i and column j if and only if the ith H_d is contained in the jth H_k.

Theorem 3.2. *Let $h(m, l) = \binom{m}{l}^n (l!)^{n-1}$. Then $M(d, k, m; n)$ is an $h(m, d) \times h(m, k)$ d-disjunct matrix with row weight $h(m-d, k-d)$ and column weight $\binom{k}{d}$.*

Proof. It is easy to see that $h(m, l)$ is the number of all distinct H_l of G. Thus, $M(d, k, m; n)$ is an $h(m, d) \times h(m, k)$ matrix with row weight $h(m-d, k-d)$ and column weight $\binom{k}{d}$.

To show that $M(d, k, m; n)$ is d-disjunct, we recall Lemma 2.2. Consider $d+1$ distinct columns $C_{j0}, C_{j1}, \ldots, C_{jd}$ of $M(d, k, m; n)$. Since these $d+1$ columns are indexed by $d+1$ distinct H_k, for each $i \in [d]$ there exists a K_n^i of G such that $K_n^i \in C_{j0} \setminus C_{ji}$. Hence, there exists a $H'_d \subseteq C_{j0}$ which contains all K_n^is. If $|\{K_n^i : i \in [d]\}| < d$, we simply add more K_n in C_{j0} to $\{K_n^i : i \in [d]\}$ to form H'_d. Furthermore, since $H'_d \nsubseteq C_{ji}$ for all $i \in [d]$, C_{j0} has a 1 in row H'_d where all other C_{ji} contain 0. □

Obviously, when $n = 1$, $M(d, k, m; n)$ is Macula's construction. When $n \ge 2$, compared with Macula's construction,

$$\frac{h(m, d)}{h(m, k)} \bigg/ \frac{\binom{mn}{d}}{\binom{mn}{k}} = \frac{(mn-d)(mn-d-1)\cdots(mn-k+1)}{(m-d)^n(m-d-1)^n\cdots(m-k+1)^n} < 1.$$

Compared with Ngo and Du's construction,

$$\frac{h(m, d)}{h(m, k)} \bigg/ \frac{g(mn/2, d)}{g(mn/2, k)} = \frac{(mn-2d)(mn-2d-1)\cdots(mn-2k+1)}{(2m-2d)^n(2m-2d-2)^n\cdots(2m-2k+2)^n} < 1.$$

Compared with Zhao's construction,

$$\frac{h(m, d)}{h(m, k)} \bigg/ \frac{\binom{n}{d}m^d}{\binom{n}{k}m^k} = \frac{m^{k-d}}{(m-d)^n(m-d-1)^n\cdots(m-k+1)^n} < 1.$$

Thus the row to column ratio of $M(d, k, m; n)$ is much smaller than that of the disjunct matrices in [7, 9, 10].

Theorem 3.3. *Let $1 \le s \le d < k \le m$ and $e = \binom{k-s}{k-d} - 1$, Then $M(d, k, m; n)$ is s^e-disjunct.*

Proof. Let $C_{j0}, C_{j1}, \ldots, C_{js}$ be any $s+1$ distinct columns of $M(d, k, m; n)$. For each $i \in [s]$, there exist a $K_n^i \in C_{j0} \setminus C_{ji}$. Let $J = \{K_n^1, K_n^2, \ldots, K_n^s\}$. Then $|J| \le s$ and J is a subset of C_{j0}, which is not a subset of C_{ji} for each $i \in [s]$. If $|J| = j$, the number of d-subsets of C_{j0} containing J is $\binom{k-j}{d-j} = \binom{k-j}{k-d}$. Since $\binom{k-j}{k-d} \ge \binom{k-s}{k-d}$ whenever $j \le s$, the number of d-subsets of C_{j0} that are not subsets of C_{ji} is at least $\binom{k-s}{k-d}$. Therefore $M(d, k, m; n)$ is an s^e-disjunct matrix. □

An s^e-disjunct matrix is called *fully s^e-disjunct* if it is not $d^{e'}$-disjunct whenever $d > s$ or $e' > e$. D'yachkov *et al.* [3] discussed the error-correcting property of Macula's construction.

THEOREM 3.4 [3]. *Suppose that $1 \le s \le d < k < n$ and $e = e(s) = \binom{k-s}{k-d} - 1$. Then $M(d, k, n)$ is fully s^e-disjunct.*

For a binary matrix M of order $N \times T$, let $B(D)$ denote the Boolean sum of those columns indexed by elements of $D \subseteq [T]$, and let $d_H(B(D), B(D'))$ denote the Hamming distance between $B(D)$ and $B(D')$ where D and D' are two distinct subsets of $[T]$. Let

$$e_s = \min_{|D|=|D'|=s} d_H(B(D), B(D')).$$

The larger the parameter e_s, the better its error-correcting capacity.

D'yachkov *et al.* [2] gave lower bounds of e_s for a fully s^e-disjunct matrix.

THEOREM 3.5 [2]. *Let M be a fully s^e-disjunct matrix. Then $e_s \ge 2(e + 1)$.*

THEOREM 3.6. *Let $1 \le s \le d < k \le m$. Then $M(d, k, m; n)$ is a fully s^e-disjunct matrix with*

$$e = \binom{k-s}{k-d} - 1, \quad e_s = 2\binom{k-s}{k-d}.$$

PROOF. Note that the maximum size of E can be obtained in Theorem 3.3, which implies that $M(d, k, m; n)$ is fully s^e-disjunct.

By Theorem 3.5, $e_s \ge 2\binom{k-s}{k-d}$, so we only need to prove $e_s \le 2\binom{k-s}{k-d}$.

For all $i, j \in [k + 1], i \ne j$, $K_n^i \cap K_n^j = \emptyset$. Suppose that $Q = \{K_n^1, K_n^2, \ldots, K_n^k\}$ and $J = \{K_n^1, K_n^2, \ldots, K_n^{k+1}\} = \{K_1, K_2, \ldots, K_{k+1}\}$. Let

$$D_0 = \{\widehat{K_1}, \widehat{K_2}, \ldots, \widehat{K_{s-1}}, \widehat{K_{k+1}}\}, \quad D_0' = \{\widehat{K_1}, \widehat{K_2}, \ldots, \widehat{K_{s-1}}, \widehat{K_k}\},$$

where $\widehat{K_i} = J - \{K_i\}$. Then

$$\left|\left\{R \;\middle|\; R \in \binom{Q}{d},\ R \not\subseteq \widehat{K_1}, \widehat{K_2}, \ldots, \widehat{K_{s-1}}, \widehat{K_k}\right\}\right| = \binom{k-s}{d-s} = \binom{k-s}{k-d}.$$

By symmetry, we have that $d_H(B(D_0), B(D_0')) = 2\binom{k-s}{k-d}$, so $e_s \le 2\binom{k-s}{k-d}$. □

DEFINITION 3.7. Let $C_{j0}, C_{j1}, C_{j2}, \ldots, C_{jd}$ denote any $d + 1$ distinct columns of $M(d, k, m; n)$. An H_d is said to be *private for C_{j0} with respect to $C_{j1}, \ldots, C_{jd}$* if $H_d \subseteq C_{j0} \setminus \bigcup_{i \in [d]} C_{ji}$. Let $p(C_{j0}; C_{j1}, \ldots, C_{jd})$ denote the number of private H_d of C_{j0} with respect to $C_{j1}, \ldots, C_{jd}$.

LEMMA 3.8 [9]. *Given integers $m > d \ge 1$ and any labeled simple graph G with $|V(G)| = m$ and $|E(G)| = d$, then the number of vertex covers of size d (or d-covers, for short) of G is at least $d + 1$.*

https://doi.org/10.1017/S0004972711002565 Published online by Cambridge University Press

Theorem 3.9. *For any $d+1$ distinct columns $C_{j0}, C_{j1}, \ldots, C_{jd}$ of $M(d, m, m; n)$, then $p(C_{j0}; C_{j1}, \ldots, C_{jd}) \geq d+1$.*

Proof. Through the construction of $M(d, k, m; n)$, we know that when $k = m$, $|C_{j0} \setminus C_{ji}| \geq 2$ for each $i \in [d]$.

For each $i \in [d]$, choose arbitrarily $E_i \subseteq C_{j0} \setminus C_{ji}$ so that $|E_i| = 2$. Suppose that $C_{j0} = \{K_n^1, K_n^2, \ldots, K_n^m\}$ and each K_n^t, $t \in [m]$ is viewed as a vertex. Let G be the graph with $V(G) = C_{j0}$, $E(G) = \{E_1, E_2, \ldots, E_d\}$. Then G is a simple graph with m vertices and at most d edges. Also, $|E(G)| \leq d$ because the E_i are not necessarily distinct. For arbitrary i, any d-subset R of C_{j0} such that $R \cap E_i \neq \emptyset$ is a private H_d of C_{j0} with respect to $C_{j1}, \ldots, C_{jd}$. Note that R is nothing but a d-cover of G. To show that $p(C_{j0}; C_{j1}, \ldots, C_{jd}) \geq d+1$, we shall show that the number of d-covers of G is at least $d+1$. Since adding more edges into G can only decrease the number of d-covers, we can safely assume that G has exactly d edges and apply Lemma 3.8. □

So when $k = m$, $M(d, k, m; n)$ is d^e-disjunct ($e = d$). According to [9], we also have the following theorem.

Theorem 3.10. *Given integers $m > d \geq 1$:*

(i) *$M(d, m, m; n)$ is d-error-detecting and $\lfloor d/2 \rfloor$-error-correcting;*

(ii) *if the number of positives is known to be exactly d, then $M(d, m, m; n)$ is $(2d+1)$-error-detecting and d-error-correcting.*

Proof. For any $s, s' \in S(\overline{d}, n)$, $s \neq s'$, we can assume without loss of generality that there exists $C_{j0} \in s \setminus s'$. Theorem 3.9 implies that $|P(s) \oplus P(s')| \geq d+1$, hence Remark 2.3 shows (i). If the number of positives is exactly d, we need only consider $|s| = |s'| = d$; hence there exist $C_{j0} \in s \setminus s'$ and $C'_{j0} \in s' \setminus s$. This time, Theorem 3.9 implies $|P(s) \oplus P(s')| \geq 2d+2$. Again, Remark 2.3 yields (ii). □

References

[1] Y. J. Bai, T. Y. Huang and K. S. Wang, 'Error-correcting pooling designs associated with some distance-regular graphs', *Discrete Appl. Math.* **157** (2009), 3038–3045.

[2] A. G. D'yachkov, F. K. Hwang and A. J. Macula, 'A construction of pooling designs with some happy surprises', *J. Comput. Biol.* **12** (2005), 1127–1134.

[3] A. G. D'yachkov, A. J. Macula and P. A. Vilenkin, 'Nonadaptive and trivial two-stage group testing with error-correcting d-disjunct inclusion matrices', in: *Entropy, Search, Complexity*, Bolyai Society Mathematical Studies, 16 (Springer, Berlin, 2007), pp. 71–83.

[4] T. Y. Huang and C. W. Weng, 'A note on decoding of superimposed codes', *J. Comb. Optim.* **7** (2003), 381–384.

[5] T. Y. Huang and C. W. Weng, 'Pooling spaces and non-adaptive pooling designs', *Discrete Math.* **282** (2004), 163–169.

[6] W. H. Kautz and R. C. Singleton, 'Nonrandom binary superimposed codes', *IEEE Trans. Inform. Theory* **10** (1964), 363–377.

[7] A. J. Macula, 'A simple construction of d-disjunct matrices with certain constant weights', *Discrete Math.* **162** (1996), 311–312.

[8] A. J. Macula, 'Error-correcting nonadaptive group testing with d^e-disjunct matrices', *Discrete Appl. Math.* **80** (1997), 217–222.

[9] H. Q. Ngo and D. Z. Du, 'New constructions of non-adaptive and error-tolerance pooling designs', *Discrete Math.* **243** (2002), 161–170.

[10] P. Zhao, K. F. Diao and K. S. Wang, 'A generalization of Macula's disjunct matrices', *J. Comb. Optim.* (2010).

FENGLIANG JIN, Sch. Sci., Linyi University, Linyi, 276005, PR China
and
Sch. Math. Sci., Shandong Normal University, Jinan, 250014, PR China
e-mail: jflajj@163.com

HOUCHUN ZHOU, Sch. Sci., Linyi University, Linyi, 276005, PR China

JUAN XU, Sch. Sci., Linyi University, Linyi, 276005, PR China