Error-correcting nonadaptive group testing with 
d\textsuperscript{e}-disjunct matrices

Anthony J. Macula*

State University of New York at Geneseo, Department of Mathematics, Geneseo, NY 14454, USA

Received 2 April 1996; received in revised form 22 November 1996; accepted 21 April 1997

Abstract

\textit{d}-disjunct matrices constitute a basis for nonadaptive group testing (NGT) algorithms and binary \textit{d}-superimposed codes. The rows of a \textit{d}-disjunct matrix represent the tests in a NGT algorithm which identifies up to \textit{d} defects in a population. The columns of a \textit{d}-disjunct matrix represent binary \textit{d}-superimposed codewords. A \textit{d}-disjunct matrix \(\mu\) is called \(d^e\)-disjunct if given any \(d + 1\) columns of \(\mu\) with one designated, there are \(e + 1\) rows with a 1 in the designated column and a 0 in each of the other \(d\) columns. \(d^e\)-disjunct matrices form a basis for \(e\) error-correcting NGT algorithms. In this paper, we construct \(d^e\)-disjunct matrices. In so doing, we simultaneously construct \(e\) error-correcting binary \textit{d}-superimposed codes. The results of this paper can be used to construct pooling designs for the screening recombinant DNA libraries. Such screenings are a major component of the Human Genome Project.

1. Introduction

Let \(n\) and \(z\) be positive integers and let \([z]\) denote \(\{1, 2, \ldots, z\}\). Given set \(S\), \(|S|\) denotes its cardinality. We call a subset of \([n]\) with cardinality \(k\) a \textit{k-set}, and we call a binary vector with \(n\) components an \textit{n-vector}. For \(k \in [n]\), \(\binom{k}{z}\) denotes the family of \(k\)-sets of \([n]\), and for \(d \leq z\), \([\leq d]\) denotes the family of subsets of \([z]\) with cardinality at most \(d\). For an \(n\)-vector \(x\), \(x_i\) denotes the \(i\)th component of \(x\). The \textit{boolean sum} of two \(n\)-vectors \(x\) and \(y\) is defined coordinate-wise using the rule \(x_i \lor y_i = 0\) if and only if \(x_i\) and \(y_i\) are both zero; otherwise \(x_i \lor y_i = 1\). A \(0, 1\) matrix with \(z\) columns is called a \textit{matrix on} \([z]\). We let \(r_i(\mu)\) and \(c_j(\mu)\) denote the \(i\)th row and \(j\)th column of \(\mu\), respectively, and often just write \(r_i\) and \(c_j\) when the context is clear.

*E-max: macula@geneseo.edu.
2. \textit{d}$^e$-disjunct matrices

**Definition 1.** For a matrix \( \mu \) on \([z]\), \( B_d(\mu) \) is the set of all boolean sums of the form \( \bigvee_{j \in D} c_j(\mu) \) where \( D \in \binom{[z]}{d} \).

**Definition 2.** A matrix \( \mu \) on \([z]\) is \textit{d}-disjunct if each \( o(\mu) \in B_d(\mu) \), with \( o(\mu) = \bigvee_{j \in D} c_j \), has the property that for each \( c_j \), with \( j \notin D \), it follows that \( c_j \lor o(\mu) \neq o(\mu) \).

If \( \mu \) is \( d \)-disjunct, \( o(\mu) \in B_d(\mu) \), and \( D = \{ c_j : c_j \leq o(\mu) \} \), then \( o(\mu) = \bigvee D \). In other words, \( o(\mu) \) is the boolean sum of the columns in \( \mu \) that are below it.

**Definition 3.** A matrix \( \mu \) is \textit{d}$^e$-disjunct if given any \( d + 1 \) columns of \( \mu \) with one designated, there are \( e + 1 \) rows with a 1 in the designated column and a 0 in each of the other \( d \) columns.

**Proposition 1** (Du and Hwang [3]). A matrix \( \mu \) is \( d \)-disjunct if and only if it is \( d^0 \)-disjunct. That is, given any \( d + 1 \) columns of \( \mu \) with one designated, there is a row with a 1 in the designated column and a 0 in each of the other \( d \) columns.

**Proposition 2.** Let \( \mu \) be a \( d^e \)-disjunct matrix on \([z]\). Let \( o(\mu) \in B_d(\mu) \) and let \( D = \{ c_j(\mu) : c_j(\mu) \leq o(\mu) \} \). Let \( |E| \leq e \) and let \( \mu_E \) be the submatrix of \( \mu \) that arises by deleting the rows \( r_i(\mu) \) for each \( i \in E \). Let \( o(\mu_E) \) be the subvector of \( o(\mu) \) that arises by deleting from \( o(\mu) \) the entries in \( E \), and let \( D_E = \{ c_j(\mu_E) : c_j(\mu_E) \leq o(\mu_E) \} \). Then

(a) \( \mu_E \) is a \( d \)-disjunct matrix on \([z]\).
(b) \( c_j(\mu_E) \in D_E \) if and only if \( c_j(\mu) \in D \).

**Proof.** (a) Apply Proposition 1.
(b) Clearly, if \( c_j(\mu) \in D \), then \( c_j(\mu_E) \in D_E \). On the other hand, without loss of generality, suppose \( c_0(\mu_E) \in D_E \) and \( c_0(\mu) \notin D \). Since \( o(\mu) \in B_d(\mu) \), without loss of generality, \( D = \{ c_1(\mu), \ldots, c_d(\mu) \} \) is the unique set of at most \( d \) distinct columns of \( \mu \) with boolean sum \( o(\mu) \). Since \( c_0(\mu_E) \in D_E \), there are at most \( e \) indices, \( i \in E \), with \( c_i, o(\mu) = 1 \) and \( o_i(\mu) = 0 \). In other words, there are at most \( e \) indices where \( c_i, o(\mu) = 1 \) and each of the entries \( c_i, 1(\mu), \ldots, c_i, e(\mu) \) are zero. This contradicts the assumption that \( \mu \) is a \( d^e \)-disjunct matrix.

For \( n \)-vectors \( x \) and \( y \), the Hamming distance \( H(x, y) \) denotes the number of corresponding components of \( x \) and \( y \) that are different. For a set of vectors \( \mathcal{V} \), let \( H(\mathcal{V}) \) denote the minimum Hamming distance between any pair of vectors in \( \mathcal{V} \). Proposition 3 follows easily from Definition 3.

**Proposition 3.** If a matrix \( \mu \) is \( d^e \)-disjunct, then \( H(B_d(\mu)) \geq 2e + 2 \).
3. Error-correcting nonadaptive group testing

Suppose we have a finite ground set containing elements which can be characterized as being either good or defective. We refer to the collection of defective elements as the defective subset and we denote it by $D$. In the abstract group testing problem, $D$ must be identified by performing 0, 1 tests on subsets of the ground set. A test result is 1 if a defect is present in a tested subset; the test result is 0 otherwise. In nonadaptive group testing (NGT), there is the added difficulty of deciding exactly which subsets to test before any testing occurs. Often, a NGT algorithm is referred to as a one stage algorithm. Every parallel algorithm is nonadaptive.

We identify a ground set (population) of cardinality $z$ with the set of columns $\{c_j(\mu)\}_{j \in [z]}$ of a matrix $\mu$ on $[z]$. Then a row of $\mu$ determines a subset of the ground set in the obvious way. That is, $c_j(\mu)$ is in the subset of the ground set determined by $r_i(\mu)$ if and only if $\mu_{i,j} = 1$. We identify a row of a matrix $\mu$ on $[z]$ with the subset of the column of $\mu$ that it determines.

If we have at most $d$ defects in our ground set and a reliable testing procedure that will detect the presence of a defect in a tested subset, then a $d$-disjunct matrix $\mu$ provides the basis for a NGT algorithm that identifies the defective subset. By testing each row of $\mu$, we define an output vector $o(\mu)$, where $o_i(\mu) = 1$ if a defect is present in $r_i(\mu)$ and 0 if not. Since there are at most $d$ defects in our ground set, it follows that $o(\mu)$ is in $B_d(\mu)$. If $o(\mu)$ is the zero vector, then $D = \emptyset$. If not, then because $\mu$ is $d$-disjunct, it follows that $D = \{c_j(\mu): c_j(\mu) \leq o(\mu)\}$.

If at most $e$ tests are unreliable, then a $d^e$-disjunct matrix can be used. A NGT algorithm based on a $d^e$-disjunct matrix is $e$ error-correcting because when any $e$ rows are deleted, the resulting submatrix is $d$-disjunct. In Propositions 4 below, suppose the following:

1. $\mu$ is a $d^e$-disjunct matrix of $[z]$, $D$ is the defective subset, and $|D| \leq d$.
2. The testing procedure defined above has been carried out to yield an output vector $o'(\mu)$.
3. No more than $e$ testing errors have occurred and $o(\mu)$ is the correct output vector.
4. Let $E = \{i: o'_i(\mu) \neq o_i(\mu)\}$ and let $\mu_E$ and $o(\mu_E)$ be as in Proposition 2. Let $o'(\mu_E)$ be the subvector of $o'(\mu)$ that arises by deleting from $o'(\mu)$ the entries in $E$.
5. Let $D' = \{c_j(\mu): c_j(\mu) \leq o'(\mu)\}$ and $D'_E = \{c_j(\mu_E): c_j(\mu_E) \leq o(\mu_E)\}$.
6. A 0 (1) test result misrecorded as a 1 (0) is called a positive (negative) error.
7. For $Z \subset \{i: o'_i(\mu) = 0\}$, define $o'_Z(\mu)$ by $o'_Z(\mu) = o'_i(\mu)$ if and only if $i \notin Z$.

Proposition 4. (a) $D' \subset D$. Moreover, if all the errors are positive, then $D = D'$.

(b) If $|D'| = d$, then $D = D'$.

(c) $H(\sqrt{D'}, o'(\mu)) \leq e + 1$ if and only if $D = D'$.

(d) If $H(\sqrt{D'}, o') \geq e + 2$, then there is a $Z$ with $|Z| \leq e$ such that $D = \{c_j(\mu): c_j(\mu) \leq o'_Z(\mu)\}$.
Proof. (a) Since $o(\mu E) = o'(\mu E)$, then if $c_j(\mu) \in D'$, it follows that $c_j(\mu) \in D$. To get $c_j(\mu) \in D$, apply Proposition 2(b). If all the errors are positive, then $o(\mu) \leq o'(\mu)$. Hence $D \subseteq D'$.

(b) Apply part (a) here.

(c) Since by assumption $H(o(\mu), o'(\mu)) < e$, it follows that $H(\sqrt{D'}, o'(\mu)) \leq 2e + 1$. Apply Proposition 3.

(d) By part (a) here, it suffices to correct the negative errors. $\square$

When $H(\sqrt{D'}, o') \geq e + 2$, one can identify the defective subset by applying Proposition 4(d) and searching over all possible $o'(\mu)$.

4. The main results

Definition 4. Let $\mathcal{X}$ be a family of $k$-sets of $[n]$. For $1 \leq d < k < n$, define the $0, 1$ matrix $\delta(n, d, \mathcal{X})$ by letting its rows and columns be, respectively, represented by the members of $\binom{n}{d}$ and $\mathcal{X}$ in the following way: For $D \in \binom{n}{d}$ and $K \in \mathcal{X}$ the matrix $\delta(n, d, \mathcal{X})$ has a 1 in its $(D, K)$th entry if and only if $D \subseteq K$. If $\mathcal{X} = \binom{n}{k}$, then we write $\delta(n, d, k)$ for $\delta(n, d, \mathcal{X})$.

Definition 5. The complement $\mu^c$ of a matrix $\mu$ on $[2]$ is the matrix that results when one interchanges the 0's and 1's in $\mu$. We defined the matrix $\delta^*(n, d, \mathcal{X})$ as that which results by row augmenting the matrix $\delta(n, d, \mathcal{X})$ with $\delta^c(n, 1, \mathcal{X})$. See Fig. 1.

Theorem 1. $\delta^*(n, d, k)$ is $d^1$-disjunct.

Proof. Let $c_0, c_1, \ldots, c_d$ be $d + 1$ distinct columns of $\delta(n, d, k)$ with $c_0$ being distinguished. For the columns $c_0, c_1, \ldots, c_d$, there are distinct corresponding $k$-sets $K_0, K_1, \ldots, K_d$ of $[n]$. Now for each $i \in [d]$, there is an $x_i \in K_0 \setminus K_i$, so the set $\{x_i\}_{i \in [d]}$ is contained in at least one $d$-set $D$ of $[n]$ with $D \subseteq K_0$ and $D \neq K_i$. Thus the row of $\delta(n, d, k)$ that corresponds to $D$ has a 1 in column $c_0$ and a 0 in each of the columns $c_1, \ldots, c_d$. Now if for some $i \in [d]$, we have that $|K_0 \setminus K_i| \geq 2$, or, if for distinct $i, j \in [d]$, we have that $K_0 \setminus K_i \cap K_0 \setminus K_j \neq \emptyset$, then it is easy to see that there is another $d$-set $D'$

\[\begin{array}{c|c|c}
\text{n rows} & \delta(n,d,k) & \delta^c(n,1,k) \\
\hline
\text{d columns} & \text{n}\binom{n}{k}\end{array}\]

Fig. 1. $\delta^*(n, d, k)$. 
with $D \neq D'$ and $D' \subset K_0$ and $D' \neq K_i$ for all $i \in [d]$. So, suppose neither of these two conditions hold. Then, for each $i \in [d]$, $|K_0 \setminus K_i| = 1$ and the family $\{K_0 \setminus K_i\}_{i \in [d]}$ is pairwise disjoint. From this, it follows that there is a $y \in (\bigcap_{i \in [d]} K_i) \setminus K_0$: Consider $K_0$ and two other distinct $K_i$ and $K_j$. Since $K_0 \setminus K_i = \{x_i\}$ and $K_0 \setminus K_j = \{x_j\}$, it follows that $K_i \setminus K_0 = \{y_i\}$ and $K_j \setminus K_0 = \{y_j\}$. A moment's reflection reveals that since $x_i \neq x_j$, then $y_i = y_j$. We let $r^*_{y_i}$ denote the row of $\delta^*(n, d, k)$ that is the complement of the row corresponding to $\{y_i\}$ in $\delta(n, 1, k)$. Then $r^*_{y_i}$ is another row of $\delta^*(n, d, k)$ with a 1 in the designated column $c_0$ and a 0 in each of the other $d$ columns $c_1, \ldots, c_d$.

**Lemma 1.** Let $\mathcal{K}$ be a family of $k$-sets in $[n]$ with $|K \setminus K'| \geq t$ for all $K$ and $K'$ in $\mathcal{K}$. Let $d \geq 1$ with $t \geq 1 + t/(k - d)$ and set $\alpha_d = \min(t^d, k - d)$. Then given $d + 1$ $k$-sets $\{K_i\}_{i=0}^d \subset \mathcal{K}$, there are $\alpha_d$ $d$-sets in $[n]$, $\{D_j\}_{j=1}^{\alpha_d}$, such that each $D_j$ is contained in $K_0$ and no $D_j$ is contained in $K_i$ for $1 \leq i \leq d$.

**Proof.** We proceed by induction. The result is clearly true when $d = 1$. Now suppose $t \geq 1 + t/(k - d)$ and the result is true for $d - 1$. Consider the $k$-set $\{K_i\}_{i=0}^d \subset \mathcal{K}$. By the inductive hypothesis, there are $\alpha_{d-1}$ $(d - 1)$-sets in $[n]$, $\{H_j\}_{j=1}^{\alpha_{d-1}}$, such that each $H_j$ is contained in $K_0$ and no $H_j$ is contained in $K_i$ with $1 \leq i \leq d - 1$. Let $\{x_i\}_{i=1}^{\alpha_{d-1}} \subset K_0 \setminus K_d$. If for each $j$ with $1 \leq j \leq \alpha_{d-1}$, we have $H_j \subset K_d$, then the $t \alpha_{d-1}$ $d$-sets $\{x_i\} \cup H_j: 1 \leq s \leq t$ and $1 \leq j \leq \alpha_{d-1}$ have the property that each $\{x_i\} \cup H_j$ is contained in $K_0$ and no $\{x_i\} \cup H_j$ is contained in $K_i$ for $1 \leq i \leq d$. It is straightforward to verify that $t \alpha_{d-1} \geq \alpha_d$. If there is an $H_j$ not contained $K_d$, then each of the $k - d + 1$ $d$-sets $\{x_i\} \cup H_j: x \in K_0 \setminus H_j$ have the desired property. Clearly $k - d + 1 \geq \alpha_d$.

**Theorem 2.** Let $\mathcal{K}$ be a family of subsets of $k$-sets of $[n]$ and $\alpha_d = \min(t^d, k - d)$. Let $H(\mathcal{K})$ denote the minimum Hamming distance (cardinality of the symmetric difference) between any pair of $k$-sets in $\mathcal{K}$. If $H(\mathcal{K}) \geq 2t$, then $\delta(n, d, \mathcal{K})$ is $d^{\alpha_d-1}$-disjunct.

**Proof.** This follows directly from Lemma 1 and Definitions 3 and 4.

**5. Some examples**

**Example 1.** $\delta^*(25, 2, 12)$ forms the basis for a 1-error correcting NGT algorithm that identifies up to two defects in a ground set of size 5.2 million using 300 tests. In the same ground set, $\delta^*(25, 3, 12)$ is 1-error correcting NGT and will identify up three defects using 2300 tests. Stirling's approximation for $n!$ implies that the column to row ratio for $\delta^*(n, d, 3)$ approaches $2^{n + 1} (\frac{2}{\pi})^{d+1/2}e^{-d}$ for large $n$.

**Example 2.** Using constant weight error-correcting codes, Theorem 2 provides NGT algorithms with considerable error correction. From [2], with $n = 25$ and $k = 12$, there is a family of 12-sets $\mathcal{F}$ with $H(\mathcal{F}) = 4$ and $|\mathcal{F}| = 227168$. So, in a ground set of size 227168, $\delta(25, 2, \mathcal{F})$ and $\delta(25, 3, \mathcal{F})$ form the basis for 3-error correcting NGT.
algorithms that identifies up to 2 and 3 defects using 300 and 2300 tests respectively. With \( n = 26 \) and \( k = 13 \), there is a family of 13-sets \( \mathcal{L} \) with \( H(\mathcal{X}) = 6 \) and \( |\mathcal{L}| = 15,031 \). Thus, in a ground set of size 15,031, \( \delta(26, 2, \mathcal{L}) \) and \( \delta(26, 3, \mathcal{L}) \) form the basis 9-error correcting NGT algorithms that identify up to 2 and 3 defects using 325 and 2600 tests, respectively.

In regard to purely error-correcting nonadaptive algorithms, the results above probably have direct practical applications only for the two and three defect cases. However, the above ideas can be used to get good, practical, and error-correcting two-stage algorithms when there are more than three defects. See [7]. Also, the results of this paper can be used to construct pooling designs for the screening recombinant DNA libraries. Such screenings are a major component of the Human Genome Project, see [1, 4] and the references therein.

References