Note

On Macula’s error-correcting pool designs☆

F.K. Hwang

Department of Applied Mathematics, National Chiao Tung University, Hsin-chu 30050, Taiwan, ROC

Received 11 December 2001; received in revised form 15 November 2002; accepted 3 December 2002

Abstract

We show that Macula’s claim of a Hamming distance 4 between any two candidate sets of positive clones in his pool design is incorrect. However, a previous proof of his on a weaker result (with a condition on design parameters) is correct. We also show that the condition is sharp and the distance 4 result is also sharp for arbitrary parameter values.

© 2003 Elsevier Science B.V. All rights reserved.

Keywords: Pooling designs; Group testing; Error-correcting; Disjunct matrix

1. Introduction

A clone library stores clones which are subsequence of a particular DNA sequence. Often, one needs to know which clones contain a given probe, a specified DNA subsequence of interest. We will call a clone positive if it contains the probe, and negative if not. It would be time-consuming and costly if we have to assay the clones one by one. Since typically the number of positive clones is small, one can pool a subset of clones together for an assay. The assay outcome is negative if all clones in the pool are negative, and is positive otherwise. A pool design is a \( 0 - 1 \) matrix where columns represent clones, rows represent pools and an 1-entry in cell \((i,j)\) signifies that clone \(j\) is in pool \(i\). The goal of a pool design is to identify the positive clones from the negative clones as much as possible with a minimum number of pools.

For a binary matrix with \(t\) rows, we can view each column as a subset of the set \(\{1, \ldots, t\}\) in terms of the positions of the 1-entries. Such a matrix is called \(d\)-disjunct if no column is contained in the union of any other \(d\) columns. It is well known [1] that

☆ Research partially supported by the Republic of China NSC grant 90-2115-M-009-029.
E-mail address: fhwang@math.nctu.edu.tw (F.K. Hwang).
a $d$-disjunct matrix can identify all positive clones as long as the number $p$ of positive clones satisfies $p \leq d$. Recently, Macula [3] introduced the notion of $d^e$-disjunct if any column has at least $e + 1$ 1-entries not in the union of any other $d$ columns. Another relevant notion is the Hamming distance $H(M)$ of a $d$-disjunct matrix $M$ which is defined to be the minimum number of bit disagreement between a union of $u$ columns and a union of $v$ columns, $u \leq v \leq d$.

Macula [2] gave a construction of a $d$-disjunct matrix. Suppose there are $z$ clones to be screened. Select $n, k, d$ such that $d \leq k$ and $\binom{n}{k} \geq z$. Let $[n]$ denote the set $\{1, \ldots, n\}$ and $\binom{[n]}{k}$ the set of all $k$-subsets of $n$. Randomly select $z$ members of $\binom{[n]}{k}$ to label the clones (columns), and label the rows by the set $\binom{[n]}{d}$ (so there are $\binom{n}{d}$ rows). The design $\delta_z(n, d, k)$ has an 1-entry in cell $(i, j)$ if and only if the label of row $i$ is contained in the label of column $j$. Macula proved that $\delta_z(n, d, k)$ is $d$-disjunct.

Macula [3] also considered the enhanced matrix $\delta^*_z(n, d, k)$ which is obtained from $\delta_z(n, d, k)$ by adding $n$ additional pools labeled $\{\bar{1}, \bar{2}, \ldots, \bar{n}\}$, where $\bar{i}$ contains all clones whose labels do not contain $i$. He claimed that $H(\delta^*_z(n, d, k)) \geq 4$ (hence 1-error-correcting) by proving

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

We will show that this claim is wrong on several counts. Nevertheless, a previous weaker claim of Macula as reported by Du and Hwang [1] remains correct:

**Theorem 2.** Suppose $k - d \geq 3$. Then $H(\delta^*_z(n, d, k)) \geq 4$.

Further, we show that both the condition $k - d \geq 3$ and the result of distance 4 are sharp.

**2. The main result**

We first give a counter-example against Theorem 1.

**Example 1.** $\delta^*_z(5, 2, 3)$ containing three columns $C_0 = \{1, 2, 3\}$, $C_1 = \{1, 2, 4\}$, $C_2 = \{1, 3, 5\}$. It is easily verified that the only 1-entry in $C_0$ but not in the union of $C_1$ and $C_2$ is the row with label (2, 3). Hence $\delta^*_z(5, 2, 3)$ is not $d^1$-disjunct.

The problem in the proof of Theorem 1 lies in the statement that let $C_0, C_1, \ldots, C_d$ be $d + 1$ distinct columns and $|C_0 \setminus C_i| = 1$ for $1 \leq i \leq d$, then $C_0 \setminus C_i \neq C_0 \setminus C_j$ implies $C_i \setminus C_0 = C_j \setminus C_0$. The above example shows that the implication is not realized since $C_1 \setminus C_0 = 4 \neq C_3 \setminus C_0 = 5$.

Example 1 can be extended to general $d, k$ with $k \geq d$. Let

\[
C_i = [k + 1] \setminus \{k + 1 - i\}, \quad 0 \leq i \leq d - 1,
\]

\[
C_d = [k + 2] \setminus \{k - d + 1, k + 1\}.
\]
Then the two candidate sets

\{\{k - d + 1, k - d + 2, \ldots, k\}\}

Next we argue that even though Theorem 1 were correct, it would not be enough to substantiate the claim that \(H(\delta^*_z(n, d, k)) \geq 4\). This is because the two candidate sets of positive clones can differ only in one column \(C\). Then the Hamming distance between those two sets is simply the number of 1-entries in \(C\) but not in the union of the other columns, which is only guaranteed to be 2 by Theorem 1. Note that \(d^1\)-disjunct would imply \(H(\delta^*_z(n, d, k)) \leq 4\) if \(d\) is the exact number of positive clones, not just an upper bound.

In a different sense, the \(d^1\)-disjunctness is too strong a property to prove a Hamming distance 4. For example, one column in one candidate set may contribute only distance 1, while the other candidate set contributes distance 3 to compensate. The two sets have Hamming distance 4, but do not satisfy \(d^1\)-disjunctness. Note that the counter-example given at the beginning of this section is not a counter-example against Theorem 2 since it is easily verified that any two candidate sets of cardinality \(d\leq 2\) have Hamming distance at least 4. A formal proof of Theorem 2 can be found in \[1\].

Can the condition \(k - d \geq 3\) in Theorem 2 be eliminated (as in Theorem 1) or at least weakened? The following example shows that it cannot.

**Example 2.** \(\delta^*_z(7,3,5)\) containing columns \(C_1 = \{1,2,3,4,5\}, C_2 = \{1,2,3,4,6\}\) and \(C_3 = \{1,2,3,5,7\}\). Consider the two candidate sets \(\{C_1, C_2, C_3\}\) and \(\{C_2, C_3\}\). It is easily verified that they differ only in three rows with labels \(\{1,4,6\}, \{2,4,6\}, \{3,4,6\}\).

We now expand the example to arbitrary \(k\) with \(d = k - 2\) and \(d \geq 3\).

Let \(n \geq k + 2\), then \(\delta^*_z(n, k - 2, k)\) contains \(k - 2\) columns

\[C_i = [k + 1] \setminus \{k + 2 - i\}, \quad 0 \leq i \leq k - 3, \quad \text{and} \]

\[C_{k-2} = [k + 2] \setminus \{4, k + 1\}.
\]

Then the two candidate sets \(\{C_0, C_1, \ldots, C_{k-3}\}\) and \(\{C_1, \ldots, C_{k-3}\}\) differ only in rows with labels \(\{1,4,5,\ldots,k\}, \{2,4,5,\ldots,k\}\) and \(\{3,4,5,\ldots,k\}\).

Examples for \(k - d < 2\) are even easier to construct and omitted here.

Next we show that regardless of how large is \(k - d\), the guaranteed Hamming distance remains at 4.

**Example 3.** \(\delta^*_z(n,2,k)\) (where \(n \geq k + 1\)) containing three columns \(C_1 = \{1,\ldots,k\}, C_2 = \{1,\ldots,k - 1, k + 1\}, C_3 = \{1,\ldots,k - 2, k, k + 1\}\). Consider two candidate sets \(\{C_1, C_2\}\) and \(\{C_2, C_3\}\). It is easily verified that the only four different rows are those labeled by \(\{k - 1, k\}, \{k, k + 1\}, \{k - 1\}\) and \(\{k + 1\}\).

Again, Example 3 can be extended to general \(d\). Let

\[C_i = [k + 1] \setminus \{k + 2 - i\}, \quad 1 \leq i \leq d + 1.
\]

Then the two candidate sets \(\{C_1, \ldots, C_d\}\) and \(\{C_2, \ldots, C_{d+1}\}\) differ only in the four rows with labels \(\{k - d + 1, k - d + 2, \ldots, k\}, \{k - d + 2, k - d + 3, \ldots, k + 1\},\)

\(\{k - d + 2\}\) and \(\{k + 1\}\).
A referee reminds us that a $d^e$-disjunct matrix can correct $e$ errors. The decoding procedure is to take a subset $E$ of rows, and change all outcomes in these rows. Do this for all $E$ with $|E| \leq e$. Let $V$ denote the outcome vector before change, and $V_E \equiv V \cup E$ is the outcome vector after change. Then a column $C$ is positive if and only if there exists an $E$ such that $V_E$ contains $C$. To see this, note that when $E$ is the set of errors, then the outcome vector is corrected back to the errorless state in which $C$ only appears in rows with positive outcomes. On the other hand, if $C$ is negative, then the $d^e$-disjunctness guarantees that $C$ has at least $e + 1$ rows not in $V_E$, and at most $e$ of them are in $E$, hence $C$ has a row not in $V_E$.

Acknowledgements

The author thanks Y.C. Liu for providing extensions to the examples, and a referee for providing the above paragraph.

References