Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

The Construction of Disjunct Matrix for Non-Adaptive Group Testing

To cite this article: S Zahidah and A Barra 2018 *J. Phys.: Conf. Ser.* **1028** 012116

IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012116 doi:10.1088/1742-6596/1028/1/012116

The Construction of Disjunct Matrix for Non-Adaptive Group Testing

S Zahidah[1] and A Barra[2]

[1]Department of Mathematics, Universitas Airlangga, Surabaya, 60115, Indonesia
[2] Department of Mathematics, Institut Teknologi Bandung, Bandung, 40132, Indonesia

\siti.zahidah@fst.unair.ac.id

Abstract. Historically, group testing theory related to the testing of blood samples to identify a disease. Based on the algorithm, there are two types of group testing, Adaptive Group Testing (AGT) and Non-Adaptive Group Testing (NAGT). NAGT algorithm can be represented by a binary matrix $M = m_{ij}$, where columns are labeled by items and rows by tests (blocks). Criteria of matrix is $m_{ij} = 1$ if test *i* contains item j and the other $m_{ij} = 0$. On the other hand, the test results of each block are represented by a column vector, called outcome vector. Based on these representations, the problem of group testing can be viewed as finding representation matrix M which satisfies the equation $Mx = y$, where *y* is an outcome vector and *x* are tested samples. If there are *d* positive samples of *n* samples then we say *d*-Combinatorial Group Testing, abbreviated by *d*-CGT. This paper presents two constructions of disjunct matrix. The first construction based on generating of binary matrices and the second construction using modular equation. Furthermore, from the construction will be modified such that the new construction can be identified more than *d* positive samples.

1. Introduction

In the beginning, group testing emerged because of the problem of detecting a disease for many samples with a minimum number of tests. The most popular problem came from R. Dorfman [1], the problem of detecting syphilitic antigen of the United States military during World War II. Along with the development of theory, group testing is now widely applied in many fields, such as DNA Sequences [2], Data Stream Algorithms [3], Digital Forensic [4], etc.

Mathematically, given N population with each sample having two possible test results, i.e. positive (+) or negative (-) and grouping into some blocks, then we can define the test function as follows :

$$T :\ 2^N \to \{\pm\}$$

where $T(X)$ is positive if X contains at least 1 positive sample and $T(X)$ is negative if all elements of X are negative. In this case we will find the set of $D \subseteq N$ which all its members are positive. If $|D| = d$ then it says d-Combinatorial Group Testing, abbreviated by *d-CGT* [5].

Based on the algorithm, there are two types of group testing:

1. Adaptive Group Testing (AGT), the determination of the test is done by taking into the results of previous tests.
2. Non-Adaptive Group Testing (NAGT), the determination of group (block) is done first and the test is performed only once for each group so that in once test it can identify all positive samples and negative samples.

IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012116 doi:10.1088/1742-6596/1028/1/012116

Grouping by NAGT algorithm can be represented by a binary matrix. This paper will show the construction of binary matrix which is the solution for *d*-CGT. The Authors use MATLAB for this construction.

2. Problems Definition

Chee [6] defined the set system for NAGT as follows:

Definition 2.1 Let X be a finite set. A *set system* S is a pair $(X, \mathcal{A})$, where:

- X is a set of samples
- $\mathcal{A} \subseteq 2^X$ is a set of subset of X, called blocks
- $|X|$ is the number of samples and $|\mathcal{A}|$ is the number of blocks

Definition 2.2 Let $A_1, A_2, \dots, A_t \in \mathcal{A}$ and $|\mathcal{A}| = t$, the *dual* of S is a set system $S^* = (Y, \mathscr{B})$ where:

- $Y = \{1,2,\dots,t\}$ and $\mathscr{B} = \{B_x : x \in X\}$
- $B_x = \{A \in \mathcal{A} : x \in A\}$
- $|Y| = |\mathcal{A}|$ and $|\mathscr{B}| = |X|$

Definition 2.3 A set system $(X, \mathcal{A})$ is *d-union free* if $\bigcup_{k=1}^{d} A_{i_k} \neq \bigcup_{k=1}^{d} A_{j_k}$ for all $A_{i_k}, A_{j_k} \subseteq \mathcal{A}$.

Hwang and S*ó*s [7] give solution for *d*-CGT problems by the following theorem.

Theorem 2.4 A set system $S = (X, \mathcal{A})$ is the solution for *d*-CGT problem if and only if $S^* = (Y, \mathscr{B})$ is *d-union free.*

3. Result and Discussion

3.1. Transformation of Set into Binary Vector

Let $A \subseteq [n]$. Binary vector of A is $\mathbf{vec}(A) = (a_1, a_2, \dots, a_n)$, where $a_i = \begin{cases} 0, & i \notin A \\ 1, & i \in A \end{cases}$.

Definition 3.1.1

i. Let $a, b \in \{0,1\}$. Boolean sum of a and b, denoted $a \oplus b$ is defined by:

$$a \oplus b = \begin{cases} 0, & a = b = 0 \\ 1, & \text{otherwise} \end{cases}$$

ii. Let $U, V \subseteq \{0,1\}^n$. The union of U and V is defined by:

$$U \uplus V = \{u_1 \oplus v_1, \dots, u_n \oplus v_n\}$$

iii. A $\mathbf{vec}(A)$ is said to be contained in $\mathbf{vec}(B)$, denoted $\mathbf{vec}(A) \subseteq \mathbf{vec}(B)$, if set A is a subset of set B.

iv. A set of $\mathbf{vec}(A_i)$ is *d-separable* if $\biguplus_{k=1}^{d} \mathbf{vec}(A_{i_k}) \neq \biguplus_{k=1}^{d} \mathbf{vec}\,(A_{j_k})$.

Lemma 3.1.2 Let $A_k \subseteq [n]$. $\mathbf{vec}\left(\bigcup_{k=1}^{d} A_k\right) = \biguplus_{k=1}^{d} \mathbf{vec}(A_k)$.

Proof. Let $\mathcal{X} = \mathbf{vec}\left(\bigcup_{k=1}^{d} A_k\right)$, write $\mathcal{X} = (x_1, x_2, \dots, x_n)$. Let $Y_k = \mathbf{vec}(A_k)$. If $x_i = 1$ then $i \in \bigcup_{k=1}^{d} A_k$ so $i \in A_k$ for some *k*. Consequently, the *i*-th component of Y_k is 1. If $x_i = 0$ then $i \notin \bigcup_{k=1}^{d} A_k$ so $i \notin A_k$ for all *k*. Consequently, the *i*-th component of Y_k is 0 for all *k*. It can be concluded $\mathcal{X} = \biguplus_{k=1}^{d} Y_k$.

Theorem 3.1.3 A set system $([n], \mathcal{A})$ is *d*-union free if and only if $\{\mathbf{vec}(A)\}$ is *d*-separable.

Proof. Suppose a set system $([n], \mathcal{A})$ *d*-union free, then any two collections which are contained *d* sets, satisfy $\bigcup_{k=1}^{d} A_{i_k} \neq \bigcup_{k=1}^{d} A_{j_k}$, where A_{i_k}, $A_{j_k} \in \mathcal{A}$. From here we get $\mathbf{vec}(\bigcup_{k=1}^{d} A_{i_k}) \neq \mathbf{vec}(\bigcup_{k=1}^{d} A_{j_k})$, then from lemma we get $\biguplus_{k=1}^{d} \mathbf{vec}(A_{i_k}) \neq \biguplus_{k=1}^{d} \mathbf{vec}(A_{j_k})$. It means $\{\mathbf{vec}(A)\}$ *d*-separable. Conversly, let $\{\mathbf{vec}(A)\}$ *d*-separable, then $\biguplus_{k=1}^{d} \mathbf{vec}(A_{i_k}) \neq \biguplus_{k=1}^{d} \mathbf{vec}(A_{j_k})$. From lemma we get $\mathbf{vec}(\bigcup_{k=1}^{d} A_{i_k}) \neq \mathbf{vec}(\bigcup_{k=1}^{d} A_{j_k})$ and then $\bigcup_{k=1}^{d} A_{i_k} \neq \bigcup_{k=1}^{d} A_{j_k}$, for A_{i_k}, $A_{j_k} \in \mathcal{A}$. Thus the set system $([n], \mathcal{A})$ is *d-union free*.

3.2. Representation Matrix for NAGT

The NAGT algorithm can be represented by a binary matrix $M = m_{ij}$, where columns are labeled by items and rows by tests (blocks). Criteria of matrix is $m_{ij} = 1$ if test *i* contains item *j* and the other

$m_{ij} = 0$. On the other hand, the test results of each block is represented by a column vector, called outcome vector. Based on these representations, the problem of group testing can be viewed as finding representation matrix M which satisfies the equation $Mx = y$ where y is an outcome vector and x are tested samples. Let M be a $t \times n$ binary matrix, M_j denotes the *j*-th column of matrix M.

Based on theorem 2.4 and theorem 3.1.3, the authors look for the representation matrix M whose columns are *d*-separable. The following are given two matrices that satisfy the solution of *d*-CGT.

Definition 3.2.1 A $t \times n$ binary matrix M is *d-separable* if and only if for any $S_1 \neq S_2 \subseteq [n]$ where $|S_1|, |S_2| = d$ such that $\biguplus_{i\in S_1} M_i \neq \biguplus_{j\in S_2} M_j$.

Definition 3.2.2 A $t \times n$ binary matrix M is *d-disjunct* if for any $S \subseteq [n]$ where $|S| = d$ and for any $j \notin S$ such that $M_j \nsubseteq \biguplus_{i\in S} M_i$.

Properties 3.2.3 If M is *d*-disjunct matrix then M is *d*-separable matrix.

3.3. The Construction of Disjunct Matrix

The authors have an idea from Kautz and Singleton [8] for this construction by the following their results:

Definition 3.3.1 Let M be a binay matrix.

i. The weight of the *j*-th column, denoted w_j, is the number of 1-entry on the *j*-th colum. Also defined:
$$\underline{w} = \min_j w_j \quad \text{and} \quad \overline{w} = \max_j w_j$$
where *j* covering all the columns of M.

ii. If each column of *M* has the same weight, then *M* is called a constant weight matrix.

iii. The dot product of M_i and M_j, denoted λ_{ij}, is the number of 1-entry on the same row in M_i and M_j. We also defined:
$$\overline{\lambda} = \max_{i,j} \lambda_{ij}$$

Lemma 3.3.2 Let *M* be a binary matrix. If *M* has minimum weight $\underline{w}$ and *M* has maximum dot product of any two columns $\overline{\lambda}$ then *M* is $\left\lfloor \frac{\underline{w}-1}{\overline{\lambda}} \right\rfloor$ −disjunct matrix.

Based on lemma above, the authors have an idea to construct disjunct matrix using MATLAB by generating all of the binary matrices then determine the weight and dot product for each binary matrix.

Algorithm 1. The construction of *d*-disjunct matrix

INPUT: $n, t \in \mathbb{N}$ and $n > t$

1. For each $i \in [2^{t\times n}]$ generate $t \times n$ binary matrix M_i
2. Determine $\underline{w}_i, \overline{\lambda}_i$, and $d_i = \left\lfloor \frac{\underline{w}_i - 1}{\overline{\lambda}_i} \right\rfloor$
3. If d_i max
4. Display M_i and d_i

OUTPUT: *M* is *d*-disjunct matrix

The purpose of this program is to know while there are *n* samples and will do *t* test, then the output can detect *d* positive samples. But, we need quite long time for running the program to generate all of $t \times n$ binary matrices. So the authors try to find another construction. Yaniv Erlich et al. [9] have ideas on how to group the samples as a modular equation, as illustration can be seen in figure 1. It illustrates that representation matrix of 15 samples are grouped into 9 tests and then those tests grouped again into 2 large groups (the first group consists of 4 tests and the second group consists of 5 tests) with grouping rule is sample *j* contained in the *i*-th test if:

$$j \equiv i \bmod 4 \ \text{(first group)}$$
$$j \equiv i \bmod 5 \ \text{(second group)}$$

IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012116 doi:10.1088/1742-6596/1028/1/012116

	Samples														
Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
2	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
3	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
5	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
6	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
7	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
8	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
9	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1

Figure 1. An example of pooling of 20 samples into 12 test with 2 groups

Generally, if the tests are grouped into groups $X_1, X_2, \dots, X_w$ and grouping rule for each group is sample j contained in the i-th test if $j \equiv i \bmod X_k$ where $i = 1,2,\dots,X_k$ and $k = 1,2,\dots,w$.

Referring to lemma 3.3.2, Erlich tried to find the additional requirement for representation matrix such that has $\bar{\lambda} = 1$, by adding the following two conditions:

1. $X_1, X_2, \dots, X_w \geq \sqrt{n}$.
2. $X_1, X_2, \dots, X_w$ are coprime.

In his paper, Erlich has not provided information explicitly, so that, the authors try to formalize the construction of the representation matrix in the following theorem:

Theorem 3.3.3 Let $M = [M^1, M^2, \dots, M^w]^T$ be a $t \times n$ binary matrix, where M^k is a $X_k \times n$ block matrix, $X_k \geq \sqrt{n}$ and coprime. If the construction of matrix M for each block is determined by:

$$M^k(i,j) = \begin{cases} 1 & , \quad j \equiv i \bmod X_k \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $i = 1,2,\dots,X_k$ then $\bar{\lambda}$ of matrix M is 1.

Proof. Suppose $\bar{\lambda} > 1$, let $\bar{\lambda} = 2$, then there are 2 columns, say r-th and s-th columns, and block a, b so that $M^a(i,r) = M^a(i,s) = 1$ and $M^b(j,r) = M^b(j,s) = 1$. So we get $r \equiv s \bmod X_a$ and $r \equiv s \bmod X_b$ which mean $X_a | r - s$ and $X_b | r - s$ obtained $X_a X_b | r - s$. Whereas $X_a X_b > n$ and X_a, X_b are coprime, so that $r - s = 0$, i.e., $r = s$ which shows that the r-th column and s-th column are the same column.

Based on proof of theorem 3.3.3, the terms that provided by Erlich can be replaced by $X_k X_l > n$ and $X_1, X_2, \dots, X_w$ are coprime. Next, based on theorem 3.3.3, the authors make following algortihms for another construction of disjunct matrix using MATLAB

Algorithm 2. The construction of 1-disjunct matrix

INPUT: $n, g_1, g_2 \in \mathbb{N}$ and $g_1, g_2 > \sqrt{n}$, and gcd$(g_1, g_2) = 1$

1. M is a $(g_1 + g_2) \times n$ zeros matrix
2. For $j = 1..n$ and for $i = 1..g_1$, if $j \equiv i$ mod g_1 then $M(i,j) = 1$
3. For $j = 1..n$ and for $i = g_1 + 1..g_1 + g_2$, if $j \equiv i$ mod s then $M(i,j) = 1$
4. Display M

OUTPUT: M is 1-disjunct matrix

Algorithm 3. The construction of 2-disjunct matrix

INPUT: $n, g_1, g_2, g_3 \in \mathbb{N}$ and $g_1, g_2, g_3 > \sqrt{n}$, and g_1, g_2, g_3 are coprime

1. M is a $(g_1 + g_2 + g_3) \times n$ zeros matrix
2. For $j = 1..n$ and for $i = 1..g_1$, if $j \equiv i$ mod g_1 then $M(i,j) = 1$
3. For $j = 1..n$ and for $i = g_1 + 1..g_1 + g_2$, if $j \equiv i$ mod g_2 then $M(i,j) = 1$
4. For $j = 1..n$ and for $i = g_1 + g_2 + 1..g_1 + g_2 + g_3$, if $j \equiv i$ mod g_3 then $M(i,j) = 1$
5. Display M

IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012116 doi:10.1088/1742-6596/1028/1/012116

OUTPUT: M is 2-disjunct matrix

Generally, the construction of *k*-disjunct matrix given by the following algorithm.

Algorithm 4. The construction of *k*-disjunct matrix

INPUT: $n, g_1, g_2, .., g_k \in \mathbb{N}$ and $g_1, g_2, .., g_k > \sqrt{n}$, and $g_1, g_2, .., g_k$ are coprime

1. M is a $(\sum_{i=1}^{k} g_i) \times n$ zeros matrix
2. For $l = 1..k$
3. For $j = 1..n$ and for $i = 1..g_l$, if $j \equiv i \mod g_l$ then $M(i, j) = 1$
4. Display M

OUTPUT: M is *k*-disjunct matrix

The purpose of the second program is to know minimal the number of tests while there are *n* samples and want to detect *d* positive samples. Based on the running time of program, the second program better than the first construction. The following table shows some result of the program

Table 1. The Output of M-File

Sample (*n*)	**Positive Sample (*d*)**	**Minimum Test (*t*)**
100	1	21
100	2	34
100	3	51
100	4	70
100	5	91
1000	1	65
1000	2	100
1000	3	137
1000	4	178
10.000	1	201
10.000	2	304

3.4. Algorithm for Determining Addition Weight

Prior to adding the weight into each column of matrix *M* in the theorem 3.3.3, we also should note how to preserve $\overline{\lambda}$. So, the first step is to determine the restrictions on the samples will be tested. As illustration can be seen in figure 2. With this limitation, there is a possibilty of space for candidates of adding weight in the matrix *M*, see figure 3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1				1	x			1		x		1		
2		1				1				1				1	
3			1				1				1				1
4				1				1				1			
1	1				x	1			x		1		x		
2		1					1					1			
3			1					1					1		
4				1					1					1	
5					1					1					1

Figure 2. The elimination of space for restrictions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1		x	x	1	x		x	1	x	x		1	x	x
2	x	1		x	x	1	x		x	1	x	x		1	x
3	x	x	1		x	x	1	x		x	1	x	x		1
4		x	x	1			x	1	x			1	x	x	
1	1	x	x		x	1	x		x	x	1		x	x	x
2		1	x	x		x	1	x		x	x	1		x	x
3	x		1	x	x		x	1	x		x	x	1		x
4	x	x		1	x	x		x	1	x		x	x	1	
5	x	x	x		1	x	x		x	1	x		x	x	1

Figure 3. Possible spaces that can be added weight

Mathematically, to find the amount of space that can not be added weight is the same as finding the number of intersect between matrix columns. Let s_{kl} be the space on k-th row and l-th column that can not be filled with entry 1, and let S_j be a set of spaces that are intersect with j-th column, then from figure 2 we get:

$$S_1 = \{s_{55}, s_{16}, s_{59}, s_{1(11)}, s_{5(3)}\} \text{ dan } |S_1| = 5 = \sum_{i=2}^{15} \lambda_{1i}$$

Let S be the set of all spaces that can not be added weight (marked by x) of matrix M, then $|S| = \sum_{i=1}^{t} \sum_{j \neq i} \lambda_{ij}$. Consequently, the amount of remaining space that can be added weight (space with the addition of entry 1) is $|R| = tn - wn - |S| = (t - w)n - |S|$.

Let R_{kl} be the set of removed remaining spaces while entry of r_{kl} is 1, then from figure 4 we get $R_{41} = \{s_{54}, s_{45}, s_{46}, s_{58}, s_{4(11)}, s_{1(12)}, s_{5(12)}\}$. So, to add weights to each column of matrix is similar to finding 12 sets of R_{kl} which are disjoint, where $l = 1,2,\dots,12$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1		x	x	1	x		x	1	x	x		1	x	x
2	x	1		x	x	1	x		x	1	x	x		1	x
3	x	x	1		x	x	1	x		x	1	x	x		1
4	1	x	x	1	x	x	x	1	x		x	1	x	x	
1	1	x	x	x	x	1	x	x	x	x	1	x	x	x	x
2		1	x	x		x	1	x		x	x	1		x	x
3	x		1	x	x		x	1	x		x	x	1		x
4	x	x		1	x	x		x	1	x		x	x	1	
5	x	x	x		1	x	x		x	1	x		x	x	1

Figure 4. Addition weight and removed remaining space

From the construction in theorem 3.3.3, we have a matrix M with constant weight w and $\overline{\lambda} = 1$. So, based on lemma 3.3.2, matrix M is $(w-1)$-disjunct matrix. Generally, to add weight into each column of $(w-1)-$ disjunct matrix in theorem 3.3.3 is equivalent to finding n sets of R_{kl} which are all disjoint, where $l = 1,2,\dots,n$. But, in this case, not all $(w-1)-$ disjunct matrix can be weighted so that $\overline{\lambda} = 1$.

4. Conclusion

This paper presents two constructions of disjunct matrix. The first construction based on generating of binary matrices and the second construction using modular equation. Besides the two construction, this paper also presents the algorithm for determining addition weight and get the following results:

1. A $(w-1)-$disjunct matrix can not be added weight when $\sum_{l=1}^{n} \min_k |R_{kl}| > |R|$.
2. A $(w-1)-$disjunct matrix can be added weight when $\sum_{l=1}^{n} \max_k |R_{kl}| < |R|$.

Acknowledgements

The authors are grateful to M I Utoyo and the reviewers for their suggestions and their comments. This work is supported by Faculty of Science and Technology, Universitas Airlangga.

References

[1] Dorfman R 1943 The detection of defective members of large populations *The Annal of Mathematical Statistics* **14** pp 436-440
[2] Pevzner P A and Lipshutz R J 1994 *Proceedings of the 19th International Symposium on Mathematical Foundations of Computer Science* (London:Springer-Verlag) pp 143-158
[3] Cormode G and Muthukrishnan S 2005 What's hot and what's not: tracking most frequent items dynamically *ACM Trans. Database Syst.* **30** pp 249-27
[4] Goodrich M T, Atallah M J and Tamassia R 2005 Indexing information for data forensics *Third International Conference on Applied Cryptography and Network Security (ANCS)* pp 206-221
[5] Du D Z and Hwang F K 2009 *Group Testing and Its Applications* 2nd Edition World Scientific
[6] Chee Y M 1996 Turán-Type Problem in Group Testing: Coding Theory and Cryptography *Ph.D. Dissertation* (University of Waterloo)
[7] Hwang F K and Sós V T 1987 Non-Adaptive Hypergeometric Group Testing *Studia scient. Math. Hungaria* **22** pp 257 – 263
[8] Kautz W H and Singleton R R 1964 Nonrandom Binary Superimposed Codes *IEEE Trans. Inform. Thy.* **10** pp 363 – 377
[9] Erlich Y, Chang K, Gordon A, et al. 2009 DNA Sudoku-Harnessing High-Throughput Sequencing For Multiplexed Specimen Analysis *Genom Res.* **19** pp 1243 – 1253