

On error-tolerant DNA screening[☆]

Weili Wu^{a,*}, Yaochun Huang^a, Xiao Huang^b, Yingshu Li^c

^aDepartment of Computer Science, University of Texas at Dallas, Richardson, TX 75083, USA

^b3M Center, Building 0235-03-F-08, St. Paul, MN 55144, USA

^cDepartment of Computer Science, Georgia State University, Atlanta, GA 30303, USA

Received 4 December 2003; received in revised form 25 January 2006; accepted 3 February 2006

Available online 25 April 2006

Abstract

Given n clones with some positive ones, the problem of DNA screening is to identify all positive clones with a set of tests each on a subset of clones, called a pool and the outcome is either the pool contains a positive clone or not. In this paper, we show that for a class of designs, if we apply those for samples with d positive clones to samples with at most $d - 1$ positive clones, the error-tolerant property will have an interesting improvement. We also make a remark on decoding method for k -error-correcting d -disjunct matrix. © 2006 Elsevier B.V. All rights reserved.

Keywords: DNA screening; Error tolerant; Pooling design; Group testing

1. Introduction

Given n clones with some positive ones that each contains a *probe* from a given set of probes, the problem of DNA screening is to identify all positive clones with a family of non-adaptive tests each on a subset of clones, called a *pool* and the outcome is either the pool contains a positive clone or not. For the former outcome, the pool is said to be *positive* and for the latter, the pool is said to be *negative*. Note that each test can be represented by a pool. Hence, the construction of the family of tests is called a *pooling design*.

A pooling design consisting of t pools and dealing with n clones is usually represented by a $t \times n$ binary matrix whose entry at cell (i, j) is 1 if and only if the i th pool contains the j th clone. If a pooling design works successfully for any sample of n clones with d positive ones, then its representation matrix must satisfy the property that any two unions of d columns cannot be identical, called *d -separable* matrix. Here, by means of the union of columns, we look at each column as a subset of rows (or pools). For example, the union of three binary column vectors $(1, 0, 1)^T$, $(1, 0, 0)^T$, $(1, 0, 1)^T$ is $(1, 0, 1)^T$.

Given a $t \times n$ d -separable matrix and t test outcomes on a sample of n clones with d positive ones, what is the time complexity for decoding, i.e., to determine the specific d positive clones? A naive method is to check every subset of d clones whether the test outcomes can be matched. This would take $O(tn^d)$ time. Ming Li (mentioned in [9]) showed that if this time complexity is a polynomial with respect to both n and d , then $NP=P$. Therefore, one introduced

[☆] Support in part by National Science Foundation under grant ACI-0305567.

* Corresponding author.

E-mail addresses: weiliwu@utdallas.edu (W. Wu), xchuang@mmm.com (X. Huang), yli@cs.gsu.edu (Y. Li).

d-disjunct matrix which is a binary matrix satisfying property that any column cannot be contained by any union of *d* other columns. It is easy to see that *d*-disjunct matrix must be *d*-separable. Moreover, the time complexity for above decoding problem of *d*-disjunct matrix is $O(n)$ [1].

A binary matrix is said to be (d, k) -disjunct if any column has at least $k + 1$ 1-entries not in the union of any *d* other columns. Pooling design with (d, k) -disjunct matrix is *k*-error-correcting, that is, decoding can still be done correctly even if up to *k* errors exist in test outcomes.

There exist several methods in the literature for solving the decoding problem for (d, k) -disjunct matrix. For one direction error that error occurs only in testing on negative pools, i.e., all negative outcomes are correct, Hwang [4] showed a $O(n)$ -time decoding method for up to *k* errors. For up to $k/2$ error tests, Huang and Weng [3] showed a decoding method running in time $O(kn)$. In general case that the number of errors is up to *k*, Wu et al. [9] showed a $O(nt)$ -time decoding method.

It is well-known that a *d*-separable matrix is *k*-error-correcting if and only if the Hamming distance of any two unions of *d* columns is at least $2k + 1$, where the Hamming distance of two column vectors is the number of components that they disagree with each other. For any binary matrix *M*, let us denote by $H_d(M)$ ($H_{\bar{d}}(M)$) the least Hamming distance between two unions of (at most) *d* columns. It has been known for a while that for any (d, k) -disjunct matrix, $H_d(M) \geq 2(k + 1)$ [2]. Recently, Hwang [4] noted that there exists a matrix *M* such that $H(M) \geq 4$, but is not $(d, 1)$ -disjunct. This matrix was constructed by Macula [6] who made a wrong claim that the matrix is $(d, 1)$ -disjunct.

In this paper, we will generalize Macula’s construction [6] with simplicial complexity and moreover, show that if we apply those designs for samples with *d* positive clones to samples with at most $d - 1$ positive clones, then the error-correcting feature will receive an interesting improvement. Such an improvement is special for this kind of designs and does not hold in general. We also make some remarks on the decoding problem for *k*-error-correcting *d*-disjunct matrix.

2. Construction with simplicial complex

A simplicial complex Δ is a family of subsets of a base set *X*, satisfying condition that every subset of a member in the family also belongs to the family. All elements in *X* are called *vertices*. Every member of the family is called a *face* and furthermore called a *k*-face if it contains *k* vertices. For example, 0-face is the empty set and 1-face is a vertex.

Assume $k > d \geq 1$. Let A_1, \dots, A_t be all *d*-faces and B_1, \dots, B_n all *k*-faces of a simplicial complex Δ . By extending a construction of Macula [5], Park et al. [7] defined matrix $M(\Delta, d, k)$ with row labels A_1, \dots, A_t and column labels B_1, \dots, B_n such that the entry at cell (A_i, B_j) is 1 if and only if $A_i \subset B_j$. They showed that for $1 \leq d < k$, $M(\Delta, d, k)$ is a *d*-disjunct matrix.

Now, by following Macula’s another construction [6], we define matrix $\bar{M}(\Delta, d, k)$ by adding $|X|$ rows to $M(\Delta, d, k)$ as follows: these $|X|$ rows are labeled by \bar{x} for all $x \in X$ and the entry at cell (\bar{x}, B_j) is 1 if and only if $x \notin B_j$. For simplicity, we will call rows with labels A_1, \dots, A_t in the *first part* and rows with label \bar{x} in the *second part*.

Both $M(\Delta, d, k)$ and $\bar{M}(\Delta, d, k)$ are very good for error-tolerance when apply to samples with at most $d - 1$ positive clones.

Theorem 1. For $k > d \geq 2$, $M(\Delta, d, k)$ is $(d - 1, k - d)$ -disjunct.

Proof. Consider *d* distinct columns labeled by *d* distinct *k*-faces B_0, B_1, \dots, B_{d-1} . Choose $a_1 \in B_0 \setminus B_1, \dots, a_{d-1} \in B_0 \setminus B_{d-1}$. Set $I = \{a_1, \dots, a_{d-1}\}$. Now, for any subset I' of $d - |I|$ vertices in $B_0 - I$, $I \cup I'$ is a *d*-face such that on the row with label $I \cup I'$, column B_0 has 1-entry and columns B_1, \dots, B_{d-1} have 0-entry. Therefore, column B_0 contains at least $\binom{k-|I|}{d-|I|}$ 1-entries not appearing in the union of columns B_1, \dots, B_{d-1} . Finally, since $|I| \leq d - 1$, we have $\binom{k-|I|}{d-|I|} \geq k - d + 1$. \square

Theorem 2. For $k > d \geq 1$,

$$H_{d-1}(\bar{M}(\Delta, d, k)) \geq 2 \min \left(\binom{k-d+2}{2}, 2(k-d)+1, k \right)$$

and

$$H_{d-1}(\bar{M}(\Delta, d, k)) \geq \min \left(\binom{k-d+2}{2}, 4(k-d)+2, 2k \right).$$

Proof. We first show the first inequality. Let U be a union of $d - 1$ columns with labels B_1, \dots, B_{d-1} and U' another union of $d - 1$ columns with labels B'_1, \dots, B'_{d-1} . We claim that either

(a) U contains $\min \left(\binom{k-d+2}{2}, 2(k-d)+1 \right)$ 1-entries in the first part of rows, not in U' ,

or

(b) U contains $k-d+1$ 1-entries in the first part of rows and $d-1$ 0-entries in the second part of rows, not in U' .

To show our claim. Choose $a_{i1} \in B_i \setminus B'_1, \dots, a_{i,d-1} \in B_i \setminus B'_{d-1}$. Let $I_i = \{a_{i1}, \dots, a_{i,d-1}\}$.

If $|I_i| \leq d - 2$, then we can find $\binom{k-d+2}{2} \left(\leq \binom{k-|I_i|}{d-|I_i|} \right)$ d -faces in B_i , containing I_i . They label $\binom{k-d+2}{2}$ rows in the first part satisfying (a). Next, we may assume $|I_i| = d - 1$ for all $i = 1, \dots, d - 1$.

If there exist $1 \leq i < i' \leq d - 1$ such that $I_i \neq I_{i'}$, then we can find at least $2(k-d+1) - 1$ d -faces in either B_i or $B_{i'}$ containing either I_i or $I_{i'}$. They give at least $2(k-d)+1$ rows in the first part satisfying (a).

Now, we may assume $I = I_i$ for all $i = 1, \dots, d - 1$ and $|I| = d - 1$. Then, we can find $k-d+1$ d -faces in B_1 or \dots or B_{d-1} containing I . They give $k-d+1$ rows in the first part on which all columns B'_1, \dots, B'_{d-1} have 0-entries and at least one of columns B_1, \dots, B_{d-1} has 1-entry. For each $a \in I$, on the row with label \bar{a} , all columns B_1, \dots, B_{d-1} have 0-entries and at least one of columns B'_1, \dots, B'_{d-1} has 1-entry. Therefore, (b) holds.

In case (a), U and U' disagree on at least $\min \left(\binom{k-d+2}{2}, 2(k-d)+1 \right)$ rows and in case (b), U and U' disagree on at least k rows. Note that by exchanging the roles of U and U' , we would obtain additional at least $\min \left(\binom{k-d+2}{2}, 2(k-d)+1, k \right)$ rows on which U' and U disagree each other. Therefore,

$$H_{d-1}(\bar{M}(\Delta, d, k)) \geq 2 \min \left(\binom{k-d+2}{2}, 2(k-d)+1, k \right).$$

To show the second inequality, we need to consider one more case that either U or U' is a union of at most $d - 2$ columns. In this case, we can obtain $\binom{k-d+2}{2}$ rows, in the first part, where U and U' disagree each other. However, we cannot double the number of rows by exchanging U and U' . Therefore,

$$\begin{aligned} H_{d-1}(\bar{M}(\Delta, d, k)) &\geq \min \left(H_{d-1}(\bar{M}(\Delta, d, k)), \binom{k-d+2}{2} \right) \\ &= \min \left(\binom{k-d+2}{2}, 4(k-d)+2, 2k \right). \quad \square \end{aligned}$$

Normally, a d -disjunct matrix without isolated column (i.e. no pool is singleton) is also $(d - 1, 1)$ -disjunct [1] and in general, this is best possible. For example, there exists a 12×16 2-disjunct matrix without isolated column in ([1, p. 147]) that every column has exactly three 1-entries. This matrix cannot be $(1, 2)$ -disjunct because for two columns having one 1-entry in common, one cannot have three 1-entries not in the other one.

Theorem 1 indicates that $M(\Delta, d, k)$ is really good for error-tolerance as long as $k - d$ is large and the number of positives clones does not exceed $d - 1$.

For matrix $\bar{M}(\Delta, d, k)$, Theorem 2 indicates that a similar thing happens. In fact, if we apply $\bar{M}(\Delta, d, k)$ to a sample with up to d instead of $d - 1$ positive clones, then the capability for error-tolerance would be quite small. The following theorem states such a result. Indeed, Macula's design [6] is a special case of $\bar{M}(\Delta, d, k)$ and Hwang [4] showed that the inequalities in the theorem are best possible for Macula's design. Hence, they are also best possible to $\bar{M}(\Delta, d, k)$.

Theorem 3. For $k > d \geq 1$, $H_d(\bar{M}(\Delta, d, k)) \geq 4$. For $k - d \geq 3$, $H_{\bar{d}}(\bar{M}(\Delta, d, k)) \geq 4$.

We omit the proof of this theorem since it is similar to that in [1] for Macula's design.

It is worth mentioning that all monotone graph properties are closely related to simplicial complexes. Therefore, the extension made in this paper may open lots of possibilities to do further study.

3. Remark on decoding

It has been very well studied on decoding problem for k -error-correcting d -separable matrix and (d, k) -disjunct matrix [2]. However, it still remains a problem on decoding for k -error-correcting d -disjunct matrix.

A d -disjunct matrix is said to be k -error-correcting if the Hamming distance between two unions of at most d columns is at least $2k + 1$. A question raised by Wu et al. [9] is whether the time complexity of decoding for k -error-correcting d -disjunct matrix can be a polynomial with respect to both t and n where n is the number of columns in the considered matrix. We remark that this question can be answered positively.

Theorem 4. There exists a decoding method for k -error-correcting d -disjunct matrix, running in time $O((n + t)t^k)$.

Proof. Let M be a k -error-correcting d -disjunct $t \times n$ matrix. Given outcomes from t tests on a sample of n clones with at most d positive ones, we assume that the number of errors is at most k . Our approach is as follows.

For each subset E of at most k pools, we suppose E is the set of all pools on each of which test outcome is wrong and remove all clones in negative pools not in E and all clones in positive pools in E . If the number of remaining clones is at most d , then we compute the Hamming distance between the given test outcomes and the true outcomes for the sample with all remaining clones as positives. If this Hamming distance does not exceed k , then we accept the result that all remaining ones are positive and all removed ones are all negative. Clearly, this method runs in time $O((n + t)t^k)$.

To show correctness of this method, we need to prove the following two claims:

- (1) There exists an E such that the number of remaining clones does not exceed d .
- (2) If the number of remaining clones does not exceed d , then the set of remaining clones remains the same.

Claim (1) follows from the d -disjunctness of M . Indeed, if E is the set of all pools on each of which test outcome is wrong, then all remaining clones should be positive and hence there exist at most d of them.

Claim (2) follows from the k -error-correcting property of M , by which, there exists exactly one sample with at most d positive clones such that the Hamming distance between the given test outcomes and the true outcomes on the sample does not exceed k . \square

Is there a decoding method for k -error-correcting d -disjunct matrix running in a polynomial time with respect to n , d , t and k ? The following result gives a negative answer.

Theorem 5. There does not exist a decoding method for k -error-correcting d -disjunct matrix running in a polynomial time with respect to n , d , t , and k unless $NP = P$.

To show this theorem, let us first study the decoding for k -error-correcting \bar{d} -separable matrix. A binary matrix is \bar{d} -separable if all unions of at most d columns are distinct. A \bar{d} -separable matrix M is k -error-correcting if $H_{\bar{d}}(M) \geq 2k + 1$.

The decoding for k -error-correcting \bar{d} -separable matrix is looking for a subset of at most d clones such that the number of positive pools not hit by the subset plus the number of negative pools hit by the subset does not exceed k . Thus, this decoding problem is closely related to the following problem.

TWO-SIDES-HITTING: Given two collections \mathcal{C} and \mathcal{D} of subsets of X and an positive integer d , find a subset A of at most d elements of X to minimize the total number of subsets in \mathcal{C} not hit by A and subsets in \mathcal{D} hit by A .

Indeed, \mathcal{C} is the collection of positive pools and \mathcal{D} is the collection of negative pools. What we minimize is the number of error tests for the "hitting" subset over all possible set of positive clones. The difference is that for the decoding problem, we know that the minimum value of the objective function is at most k and want to find the subset to achieve this value.

Similarly, the decoding for k -error-correcting d -disjunct matrix is closely related to a variation of TWO-SIDES-HITTING. Note that by d -disjunctness, the set of positive clones is the complement of the union of all negative pools. Thus, the minimization should be over all subsets A satisfying the following properties:

- (a) $|A| \leq d$, and
- (b) $A = X - \cup_{B \in \mathcal{C} \cup \mathcal{D}, A \cap B = \emptyset} B$.

Formally, we state this variation as follows.

TWO-SIDES-HITTING*: Given two collections \mathcal{C} and \mathcal{D} of subsets of X and an positive integer d , find a subset A , satisfying (a) and (b), to minimize the total number of subsets in \mathcal{C} not hit by A and subsets in \mathcal{D} hit by A .

Lemma 6. *If decoding for k -error-correcting d -disjunct matrix can be done in polynomial-time with respect to n, t, d , and k , then TWO-SIDES-HITTING* can be solved in polynomial time.*

Proof. The decoding for k -error-correcting d -disjunct matrix is equivalent to the problem that knowing the minimum value of TWO-SIDES-HITTING* is at most k , find a subset A satisfying (a) and (b), to achieve the minimum. If there exists an algorithm K solving this problem in polynomial time with respect to n, t, d and k , then we may solve TWO-SIDES-HITTING* by applying this algorithm, repeatedly for $k = 1, 2, \dots, t$. For each $k = 1, 2, \dots, t$, if the algorithm cannot find, in polynomial time, a subset A satisfying (a) and (b) to achieve the k -value, then restart the algorithm for next k -value, until a k -value is achieved by a subset A satisfying (a) and (b). This clearly still runs in polynomial time. \square

Next, we show that TWO-SIDES-HITTING* is NP-hard. To do so, we first study a variation of the vertex-cover problem as follows:

VERTEX-COVER: Given a graph G with n vertices and a positive integer $h, 0 < h \leq n$, determine whether G has a vertex-cover of size h .

VERTEX-COVER*: Given a graph G and two positive integers m and h , determine whether G has a vertex subset of size at most h , covering at least m edges, such that the complement of the vertex-cover has no isolated vertex.

Lemma 7. *VERTEX-COVER* is NP-complete.*

Proof. VERTEX-COVER* is clearly in NP since we can guess the subset and verify in polynomial-time. We next construct a polynomial-time reduction from the well-known NP-complete problem VERTEX-COVER to VERTEX-COVER*.

Consider a graph G with n vertices and a positive integer $h, 0 < h \leq n$. Let m be h plus the number of edges of G . We construct another graph G' from G by adding $h + 1$ new vertices v_0, v_1, \dots, v_h and connecting v_0 to all vertices of G and v_1, \dots, v_h . (Fig. 1) Next, we show that G has a vertex-cover of size h if and only if G' has a vertex subset of size h , covering at least m edges, such that the complement of the subset has no isolated vertex.

First, if G has a vertex-cover of size at most h , then this vertex cover in G' has the required property.

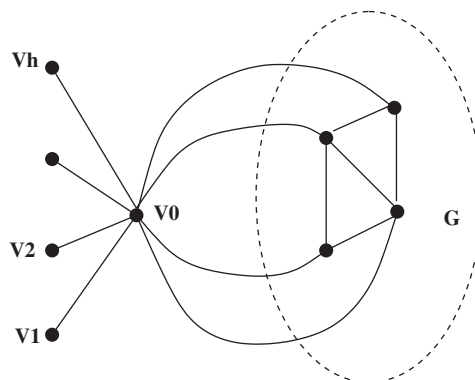


Fig. 1. Construction of G' from G .

Next, assume that G' has a vertex subset A of size at most h , covering at least m edges, such that the complement of A contains no isolated vertex. Note that v_0 cannot be in A . In fact, if v_0 is in A , then each v_i for $i = 1, \dots, h$ has to be in A ; otherwise, v_i would be isolated in the complement. However, all v_0, v_1, \dots, v_h being in A contradicts the size of A . Since v_0 is not in A and each vertex other than v_0 covers exactly one edge not in G , A covers at least $m - h$ edges in G . Hence, A covers all edges of G . Therefore, all vertices of G in A is a vertex-cover of size at most h . If this vertex-cover has size smaller than h , then we can simply add in more vertices to achieve the size h . \square

Finally, we finish the proof of Theorem 5 by proving the following lemma.

Lemma 8. TWO-SIDES-HITTING* is NP-hard.

Proof. Consider decision version of TWO-SIDES-HITTING*:

Decision version of TWO-SIDES-HITTING*: Given two collections \mathcal{C} and \mathcal{D} of subsets of X and two positive integers d and k , determine whether or not there exists a subset A , satisfying (a) and (b), and the total number of subsets in \mathcal{C} not hit by A and subsets in \mathcal{D} hit by A is at most k .

Now, we construct a polynomial-time reduction from VERTEX-COVER* to the decision version of TWO-SIDES-HITTING*. Consider an instance of VERTEX-COVER*, consisting of a graph G and two positive integers m and h . Let X be the vertex set of G and \mathcal{C} the edge set of G . Set $\mathcal{D} = \emptyset$, $k = |\mathcal{C}| - m$, and $d = h$. We show that G has a vertex subset H of size at most h , hitting at least m edges, such that its complement contains no isolated vertex if and only if X has a subset A of size at most d satisfying condition (b) and the number of subsets in \mathcal{C} not hit by A is at most k .

If G has such a vertex subset H , then set $A = H$ which is required A for X . Conversely, if X has such a subset A , then set $H = A$ and H is a required vertex subset for G . \square

4. Discussion

An *anti-chain* in a simplicial complex is a collection of faces such that no one is another's subset. It is not hard to show that all results in Section 2 hold if we use an anti-chain instead of the collection of k -faces and define k to be the smallest cardinality of a face in the anti-chain.

Since d -disjunct matrix corresponds to certain type of super imposed code, the results in Section 3 may also have impact in error-correcting code [8].

References

- [1] D.-Z. Du, F.K. Hwang, Combinatorial Group Testing and Its Applications, second ed., World Scientific, Singapore, 1999.
- [2] D.-Z. Du, F.K. Hwang, Pooling Designs: Group Testing in Biology, manuscripts, 2006.
- [3] T. Huang, C.-W. Weng, A note on decoding of superimposed codes, J. Combin. Optim. 7 (2003) 4.
- [4] F.K. Hwang, On Macula's error-correcting pooling design, Discrete Math. 2003, to appear.
- [5] A.J. Macula, A simple construction of d -disjunct matrices with certain constant weights, Discrete Math. 162 (1996) 311–312.
- [6] A.J. Macula, Error correcting nonadaptive group testing with d^e -disjunct matrices, Discrete Applied Math. 80 (1997) 217–222.
- [7] H. Park, W. Wu, Z. Liu, X. Wu, H. Zhao, DNA screening, pooling designs, and simplicial complex, J. Combin. Optim. 7 (2003) 4.
- [8] W.W. Paterson, Error Correcting Codes, MIT Press, Cambridge, MA, 1961.
- [9] W. Wu, C. Li, X. Wu, X. Huang, Decoding in pooling designs, J. Combin. Optim. 7 (2003) 4.