JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 4, 2006

Pp. 990–995

New Construction for Transversal Design

DING-ZHU DU,[1] F.K. HWANG,[2] WEILI WU,[1] and TAIEB ZNATI[3]

ABSTRACT

The study of gene functions requires the development of a DNA library of high quality through much of testing and screening. Pooling design is a mathematical tool to reduce the number of tests for DNA library screening. The transversal design is a special type of pooling design, which is good in implementation. In this paper, we present a new construction for transversal designs. We will also extend our construction to the error-tolerant case.

Key words: pooling design, transversal design, new construction.

1. INTRODUCTION

A RECENT IMPORTANT DEVELOPMENT IN BIOLOGY is the success of Human Genome Project. This project was done with a great deal of help from computer technology, which made computational biology a hot interdisciplinary research area between molecular biology, computer science, and mathematics. As the technology for obtaining sequenced genome data matures, more and more sequenced genome data are available to the scientific research community, so that the study of gene functions has become a popular research direction. Such a study is supported by a high quality DNA library which is usually obtained through much testing and screening. Therefore, the efficiency of testing and screening becomes very important. Pooling design is a mathematical tool to reduce the number of tests in DNA library screening (D'yachkov *et al.*, 2001; Farach *et al.*, 1997). For example the Life Science Division of Los Alamos National Laboratories in 1998 (Marathe *et al.*, 2000) was dealing with 220,000 clones. Testing those clones individually requires 220,000 tests. However, they used only 376 tests with pooling designs.

Pooling design is also called *nonadaptive group testing*. Given a set of n items with at most d positive ones, group testing tests subsets of items, called *pools*, instead of individual items. For example, in the above mentioned testing at Los Alamos National Laboratories, each pool contains about 5,000 clones. The outcome of a test on a pool is *positive* if the pool contains a positive item and is *negative* otherwise.

The technology of group testing was initiated from a Wasserman-type blood test in World War II. Since then, many constructions have been developed in the literature (Du and Hwang, 1999, unpublished). A group testing algorithm is said to be *nonadaptive* if all tests are arranged in a single round, that is, if no information on test outcomes is available for determining the composition of another test. A pooling design is said to be *transversal* if it can be divided into disjoint families, each of which is a partition of all items such that pools in different parts are disjoint.

[1]Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083.
[2]Department of Applied Mathematics, National Chiaotung University, Hsing Chu, Taiwan, ROC.
[3]Department of Computer Science, University of Pittsbugh, Pittsburgh, PA 15215.

In this paper, we present a new construction for transversal design. To identify n items with at most d positive ones, our construction gives a transversal design with at most $(2 + o(1))\frac{d \log n}{\log(d \log n)})^2$ tests, which is superior to all previously known transversal designs. This construction can also be easily extended to the error-tolerant case, which is an important topic in pooling designs (Ngo and Du, 2000, 2002; Hwang, 2003; Macula, 1997; Wu *et al.*, 2003, submitted).

Transversal designs are used very frequently in practice because implementation is easy and their performance is quite good. Therefore, our new construction has a significant impact in practice.

2. MATRIX REPRESENTATION OF TRANSVERSAL DESIGNS

A pooling design is usually represented by a binary matrix with rows indexed with items and columns indexed with pools. A cell (i, j) contains a 1-entry if and only if the ith pool contains the jth item. This binary matrix is called the *incidence matrix* of the represented pooling design. By treating a column as a set of row indices each intersecting the column with a 1-entry, we can talk about the union of several columns. A binary matrix is *d-separable* if every two unions from different subsets of d columns are different, is *$\bar{d}$-separable* if every two unions from different subsets of at most d columns are different, and is *d-disjunct* if no column is contained in a union of other d columns.

A transversal design has a special matrix representation with rows indexed by families and columns indexed by items; a cell (i, j) contains entry k if and only if item j belongs to the kth pool in the ith family. This matrix representation is called a *transversal matrix* of the represented transversal design.

Each $f \times n$ matrix can be seen as a transversal matrix of a transversal design as follows: Use entries on the ith row to index pools in the ith family. The pool with index k in the ith family contains the jth item if and only if cell (i, j) contains entry k in the matrix. For example, matrix

$$\begin{pmatrix} 1 & 1 & -1 & -1 \\ 2 & 3 & 2 & 3 \\ 0 & 1 & 1 & 2 \end{pmatrix}$$

represents transversal design

$$\begin{array}{lll} \{1, 2\}, & \{3, 4\}; & \\ \{1, 3\}, & \{2, 4\}; & \\ \{1\}, & \{2, 3\}, & \{4\}. \end{array}$$

In this way, each matrix represents a transversal design and each transversal design has more than one transversal matrix.

We can also extend the concept of d-separability, $\bar{d}$-separability, and d-disjunctness to the general matrix. For a general matrix, the union of d column vectors is defined to be a column vector each of whose components is the union of corresponding components of those d column vectors. A general matrix is *d_*-separable* (*$\bar{d}_*$-separable*) if all unions of (at most) d columns are different. A general matrix is d_*-disjunct if no column is contained in the union of d other columns (i.e., at least one component of the column is not contained in the corresponding component of the union). For example, the reader may verify that the 3×4 matrix in above example is 2_*-disjunct.

Theorem 1. *A transversal design is d-separable if and only if its general matrix representation is d_*-separable. A transversal design is $\bar{d}$-separable if and only if its general matrix representation is $\bar{d}_*$-separable. A transversal design is d-disjunct if and only if its general matrix representation is d_*-disjunct.*

Proof. One may transform each general matrix representation of a transversal design to its binary matrix representation by replacing each row R_i by several rows with indices each being a pair of i and an entry k of R_i and in row $\{i, k\}$, the cell $(\{i, k\}, j)$ contains a 1-entry if and only if in the general matrix representation, the cell (i, j) contains entry k. Then it is easy to verify that the resulting binary matrix is d-separable if and only if the original general matrix is d_*-separable, the resulting binary matrix

is $\bar{d}$-seperable if and only if the original general matrix is $\bar{d}_*$-separable, and the resulting binary matrix is d-disjunct if and only if the original general matrix is d_*-disjunct. ■

It may be worth mentioning that a d-separable binary matrix must be d_*-separable, but a d_*-separable binary matrix may not be d-separable. Similar relations hold between $\bar{d}$-separability and $\bar{d}_*$-separability, and between d-disjunctness and d_*-disjunctness. This is why we use d_* instead of d in the terminologies involving general matrices.

3. A NEW CONSTRUCTION

We present a new construction of transversal design in this section.

Consider a finite field $GF(q)$ of order q. Suppose k satisfies

$$n \le q^k \tag{1}$$

and

$$f = d(k-1) + 1 \le q. \tag{2}$$

We construct an $f \times n$ matrix $M(d, n, q, k)$ as follows: Its column indices are polynomials of degree k over the finite field $GF(q)$. Its row indices are f distinct elements of $GF(q)$. The cell (x, g) contains element $g(x)$ of $GF(q)$.

Theorem 2. *$M(d, n, q, k)$ is a d_*-disjunct matrix.*

Proof. Suppose $M(d, n, q, k)$ is not d_*-disjunct. Then it has a column g_0 contained in the union of other d columns $g_1, \ldots, g_d$. That is, for each row index x_i, $g_0(x_i) = g_j(x_i)$ for some j. Note that there are $d(k-1)+1$ rows. Thus, there exists a g_j $(1 \le j \le d)$ such that $g_0(x_i) = g_j(x_i)$ for at least k row indices x_i. It follows that $g_0 = g_j$, a contradiction. ■

By (1) and (2), k and q should be chosen to satisfy

$$\log_q n \le k \le \frac{q-1}{d} + 1. \tag{3}$$

There exists a positive integer k satisfying (3) if q satisfies

$$\log_q n \le \frac{q-1}{d}. \tag{4}$$

That is, it is sufficient to choose q satisfying

$$n^d \le q^{q-1}. \tag{5}$$

Let q_0 be the smallest number q satisfying (5). Then, we have the following estimation on q_0.

Lemma 3.

$$q_0 = (1 + o(1))\frac{d \log_2 n}{\log_2(d \log_2 n)}.$$

Moreover,

$$q_0 \le 1 + \frac{2d \log_2 n}{\log_2(d \log_2 n)}$$

for $n^d \ge 2^4$.

Proof. Set

$$q_1 = 1 + (1 + h(d, n)) \frac{d \log_2 n}{\log_2(d \log_2 n)},$$

where

$$h(d, n) = \frac{\log_2 \log_2(d \log_2 n)}{\log_2(d \log_2 n) - \log_2 \log_2(d \log_2 n)}.$$

Note that $h(d, n) \geq 0$. Therefore,

$$\begin{aligned}(q_1 - 1) \log_2 q_1 &> (q_1 - 1) \log_2(q_1 - 1) \\ &\geq \frac{(1 + h(d, n))d \log_2 n}{\log_2(d \log_2 n)} \cdot \log_2 \frac{(1 + h(d, n))d \log_2 n}{\log_2(d \log_2 n)} \\ &> d \log_2 n.\end{aligned}$$

That is, q_1 satisfies (5). It follows that $q_0 \leq q_1$. Note that $h(d, n) = o(1)$. Hence,

$$q_0 = (1 + o(1)) \frac{d \log_2 n}{\log_2(d \log_2 n)}.$$

Moreover, for $n^d \geq 2^4$, $d \log_2 n \geq 4$. Hence, $2^{d \log_2 n} \geq (d \log_2 n)^2$. Thus, $d \log_2 n \geq 2 \log_2(d \log_2 n)$. It follows that $h(d, n) \leq 1$. Therefore,

$$q_0 \leq 1 + \frac{2d \log_2 n}{\log_2(d \log_2 n)}$$

for $n^d \geq 2^4$. ■

We need to find a prime power q satisfying

$$q \geq q_0.$$

Then, we can choose

$$k = \lceil \log_q n \rceil.$$

For such a choice of k, we have

$$f = d(k - 1) + 1 \leq d(\lceil \log_q n \rceil - 1) + 1 \leq d(\lceil \log_{q_0} n \rceil - 1) + 1 \leq q_0.$$

Since each family contains at most q pools, the total number of tests is at most $q_0 q$.

Theorem 4. *There exist a prime power q and a positive integer k satisfying (1) and (2), such that $M(d, n, q, k)$ gives a transversal design with at most $2q_0^2$ tests.*

Proof. Set $q = 2^{\lceil \log_2 q_0 \rceil}$. Then q is a prime power satisfying $q_0 \leq q < 2q_0$). Therefore, $qq_0 < 2q_0^2$. ■

Corollary 5. *There exists a transversal design $M(d, n, q, k)$ with at most*

$$(2 + o(1)) \left(\frac{d \log_2 n}{\log_2(d \log_2 n)} \right)^2$$

tests.

There exist two previous constructions for transversal designs in the literature. The first one is the grid design (Barillot *et al.*, 1991; Hwang, 1995; Phatarfod and Sudbury, 1994). With a k-dimensional grid, the number of tests can be $O(dn^{1/k})$. The second is the Chinese remainder sieve (Eppstein *et al.*, 2004) which uses $O(\frac{(d \ln n)^2}{\ln(2d \ln n)})$ tests. Our new construction uses $O((\frac{d \log_2 n}{\log_2(d \log_2 n)})^2)$ tests, which is better than both previous ones.

Moreover, this new construction can be easily extended to the error-tolerant case. Let e be the upper bound for the number of possible errors in testing. To have an error-tolerant property, a pooling design has to meet some stronger requirement. A pooling design is *$d^{\#e}$-disjunct* if its binary representation matrix satisfies the property that every column has at least $e + 1$ 1-entries not contained in the union of some other d columns.

Lemma 6. *A transversal design is $d^{\#e}$-disjunct if and only if its general matrix representation has the property that every column has at least $e + 1$ components not contained in the union of some other d columns.*

Proof. Similar to the proof of Theorem 1. ■

Now, let us assume that q and k satisfy

$$n \le q^k \tag{6}$$

and

$$f = d(k-1) + 1 + e \le q. \tag{7}$$

We construct an $f \times n$ matrix $M(d, n, q, k, e)$ as follows: Its column indices are polynomials of degree k over the finite field $GF(q)$. Its row indices are f distict elements of $GF(q)$. The cell (x, g) contains element $g(x)$ of $GF(q)$.

Theorem 7. *$M(d, n, q, k, e)$ is a general matrix representation of $d^{\#e}$-disjunct transversal design.*

Proof. Suppose $M(d, n, q, k, e)$ is not d_*-disjunct. Then it has a column g_0 which has at least $f - e$ components contained in the union of the other d columns $g_1, \ldots, g_d$. Thus, there exists a column g_j containing at least k components of g_0. That is, for at least k row indices x_i, $g_0(x_i) = g_j(x_i)$. Therefore, $g_0 = g_j$, a contradiction. ■

By an argument similar to the above, we can also obtain the following.

Theorem 8. *By properly choosing q and k, we can obtain an $M(d, n, q, k, e)$ with at most $2q_e^2$ tests where*

$$q_e = e + (2 + o(1)) \left(\frac{2d \log_2 n}{\log_2(d \log_2 n)} \right)^2.$$

4. DISCUSSION

The coefficient 2 in Theorems 7 and 8 can be further improved if we have better knowledge of the distribution of prime powers. In fact, if Goldbach's conjecture about even numbers is true, that is, every even number is a sum of two primes, then there exists a prime between m and $2m$ for every natural number m. This means that there exist many prime powers between m and $2m$. What is the smallest constant c such that there exists a prime power between m and cm? It is unknown and possibly an interesting open problem. It may have a relation to the number of representations being a sum of two prime powers for an even number.

ACKNOWLEDGMENTS

W.W. was supported in part by NSF grant ACI-0305567 and T.Z. was supported in part by NSF grant CCF-0548895.

REFERENCES

Barillot, E., Lacroix, B., and Cohen, D. 1991. Theoretical analysis of library screening using N-dimensional pooling designs. *Nucl. Acids Res.* 19, 6241–6247.

Du, D.-Z., and Hwang, F.K. 1999. *Combinatorial Group Testing and Its Applications*, 2nd ed., World Scientific, Singapore.

Du, D.-Z., and Hwang, F.K. Unpublished. Pooling designs: Group testing in biology. Manuscript.

D'yachkov, A.G., Macula, A.J., Torney, D.C., and Vilenkin, P.A. 2001. Two models of nonadaptive group testing for designing screening experiments. *Proc. 6th Int. Workshop on Model-Orented Designs and Analysis*, 63–75.

Eppstein, D., Goodrich, M.T., and Hirschberg, D.S. 2004. Improved combinatorial group testing for real-world problem size. Manuscript.

Farach, M., Kannan, S., Knill, E., and Muthukrishnan, S. 1997. Group testing problem with sequences in experimental molecular biology. *Proc. Compression and Complexity of Sequences*, 357–367.

Hwang, F.K. 1995. An isomorphic factorization of the complete graph. *J. Combinatorial Theory* 19, 333–337.

Hwang, F.K. 2003. On Macula's error-correcting pooling design. To appear in *Disc. Math.*

Macula, A.J. 1997. Error correcting nonadaptive group testing with d^e-disjunct matrices. *Disc. Appl. Math.* 80, 217–222.

Marathe, M.V., Percus, A.G., and Torney, D.C. 2000. Combinatorial optimization in biology. Manuscript.

Ngo, H.Q., and Du, D.-Z. 2000. A survey on combinatorial group testing algorithms with applications to DNA library screening, *in Discrete Mathematical Problems with Medical Applications*, 171–182, DIMACS Series *Discrete Math. Theoret. Comput. Sci.* 55, Amer. Math. Soc., Providence, RI.

Ngo, H.Q., and Du, D.-Z. 2002. New constructions of non-adaptive and error-tolerance pooling designs. *Disc. Math.* 243, 161–170.

Phatarfod, R.M., and Sudbury, A. 1994. The use of a square array scheme in blood testing. *Statistics and Medicine* 13, 1337–1343.

Wu, W., Li, C., Huang, X., and Li, Y. Submitted. On error-tolerant DNA screening. Submitted to *Disc. Appl. Math.*

Wu, W., Li, C., Wu, X., and Huang, X. 2003. Decoding in pooling designs. *J. Combinatorial Optimization* 7(4).

Address correspondence to:
Ding-Zhu Du
Dept. of Computer Science
University of Texas at Dallas
MS EC31
2601 North Floyd Road
Richardson, TX 75083

E-mail: ding-zhu.du@utdallas.edu

This article has been cited by:

1. Haixia Guo, Jizhu Nan. 2014. Construction of error-tolerance pooling designs in symplectic spaces. *Journal of Global Optimization* **58**:2, 405-410. [CrossRef]
2. Francis Y.L. Chin, Henry C.M. Leung, S.M. Yiu. 2013. Non-adaptive complex group testing with multiple positive sets. *Theoretical Computer Science* **505**, 11-18. [CrossRef]
3. Jun Guo, Yuexuan Wang, Suogang Gao, Jiangchen Yu, Weili Wu. 2010. Constructing error-correcting pooling designs with symplectic space. *Journal of Combinatorial Optimization* **20**:4, 413-421. [CrossRef]
4. Jizhu Nan, Jun Guo. 2010. New error-correcting pooling designs associated with finite vector spaces. *Journal of Combinatorial Optimization* **20**:1, 96-100. [CrossRef]
5. Jun Guo. 2010. Pooling designs associated with unitary space and ratio efficiency comparison. *Journal of Combinatorial Optimization* **19**:4, 492-500. [CrossRef]
6. Yongxi Cheng, Ding-Zhu Du, Ker-I Ko, Guohui Lin. 2009. On the Parameterized Complexity of Pooling Design. *Journal of Computational Biology* **16**:11, 1529-1537. [Abstract] [Full Text PDF] [Full Text PDF with Links]
7. Yongxi Cheng, Ding-Zhu Du, Guohui Lin. 2009. On the upper bounds of the minimum number of rows of disjunct matrices. *Optimization Letters* **3**:2, 297-302. [CrossRef]
8. Yongxi Cheng, Ding-Zhu Du. 2008. New Constructions of One- and Two-Stage Pooling Designs. *Journal of Computational Biology* **15**:2, 195-205. [Abstract] [Full Text PDF] [Full Text PDF with Links]
9. Hung Q. Ngo. 2008. On a hyperplane arrangement problem and tighter analysis of an error-tolerant pooling design. *Journal of Combinatorial Optimization* **15**:1, 61-76. [CrossRef]
10. Ping Deng, F. K. Hwang, Weili Wu, David MacCallum, Feng Wang, Taieb Znati. 2008. Improved construction for pooling design. *Journal of Combinatorial Optimization* **15**:1, 123-126. [CrossRef]
11. Yongxi Cheng, Ding-Zhu Du. 2007. Efficient Constructions of Disjunct Matrices with Applications to DNA Library Screening. *Journal of Computational Biology* **14**:9, 1208-1216. [Abstract] [Full Text PDF] [Full Text PDF with Links]