J Comb Optim (2008) 15: 123–126
DOI 10.1007/s10878-007-9093-1

Improved construction for pooling design

Ping Deng · F.K. Hwang · Weili Wu · David MacCallum · Feng Wang · Taieb Znati

Published online: 19 July 2007

Abstract Pooling design is a mathematical tool with many applications in molecular biology, specially to reduce the number of tests for DNA library screening. In this note, we study construction of pooling design and present an improvement to a recent new construction given by Du et al. (J. Comput. Biol. 13:990–995, 2006).

Keywords Pooling design · Transversal design · Multiplication theorem · Disjunct matrix

P. Deng and W. Wu supported in part by National Science Foundation under grants CCF-0514796 and CNS-0524429. T. Znati supported in part by National Science Foundation under grant CCF-0548895.

P. Deng (✉) · W. Wu
Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083, USA
e-mail: pxd010100@utdallas.edu

W. Wu
e-mail: weiliwu@utdallas.edu

F.K. Hwang
Department of Applied Mathematics, National Chiaotung University, Hsing Chu, Taiwan

D. MacCallum · F. Wang
Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA

D. MacCallum
e-mail: fwang@cs.umn.edu

F. Wang
e-mail: dmac@cs.umn.edu

T. Znati
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15215, USA
e-mail: znati@cs.pitt.edu

1 Introduction

Given a set of n items with at most d positive ones, we study the problem of identifying all positive items with less number of tests each of which is on a subset of items, called *pools*, with two outcomes, *positive* if the pool contains a positive item, and *negative* if the pool contains no positive item. This problem is called *group testing*. A group testing algorithm is said to be *nonadaptive* if all pools are given at the beginning, that is, no information on test outcomes is available for determining the pool of another test. Nonadaptive group testing is also called *pooling design*. A pooling design is *transversal* if it can be divided into disjoint families, each of which is a partition of all items into pools in the family.

Pooling design has been found to have many applications in molecular biology. Indeed, the study of gene functions is a hot direction. Such a study is required to have DNA libraries of high quality, which usually obtained through a large amount of testing and screening. Pooling design is a mathematical tool to reduce the number of tests (Marathe et al. 2000; D'ychkov et al. 2001; Farach et al. 1997) to do those jobs.

In Du et al. (2006), Du, Hwang, Wu and Znati gave a new construction for transversal design. In this note, we propose an improvement.

A pooling design is usually represented by a binary matrix with rows indexed with pools and columns indexed with items. 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 $(d; z)$-*disjunct* if for any $d+1$ columns $C_0, C_1, \ldots, C_d$, there are at least z rows at them, C_0 contains 1-entries, all $C_1, \ldots, C - d$ contain 0-entries.

A q-nary matrix is a matrix with entries in $\{0, 1, \ldots, q-1\}$. A transversal design can be represented by a q-nary matrix with rows indexed with families and columns indexed with 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. A q-nary matrix is $(d, 1; z)$-*disjunct* if for any $d+1$ columns $C_0, C_1, \ldots, C_d$, there are at least z rows at each of which the entry of C_0 is different from the entry of C_j for $1 \le j \le d$.

A $f \times n$ q-nary matrix can be transformed into a $fq \times n$ binary matrix by replacing each i-entry with a q-dimensional column vector e_{i+1} which contains a 1-entry at the $(i+1)$th component and 0-entries at other components. The following has been known (Du et al. 2006).

Lemma 1 *A $f \times n$ q-nary matrix is $(d, 1; z)$-disjunct if and only if it can be transformed into a $fq \times n$ $(d; z)$-disjunct binary matrix.*

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

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

and

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

Let $M(d, n, q, k)$ be a $f \times n$ q-nary matrix of rows indexed with elements in $GF(q)$ and columns indexed with polynomials of degree $k-1$ over the finite field $GF(q)$; the cell (x, g) contains element $g(x)$ of $GF(q)$.

Theorem 2 *$M(d, n, q, k)$ is a $(d, 1; z)$-disjunct q-nary matrix.*

In fact, for contradiction, suppose $M(d, n, q, k)$ is not $(d, 1; z)$-disjunct. Then there are $d+1$ columns $g_0, g_1, \ldots, g_d$ such that $g_0(x) \in \{g_1(x), \ldots, g_d(x)\}$ for at least $f - z + 1$ $(= d(k-1)+1)$ x's. Thus, there exists j, $1 \le j \le d$, such that $g_0(x) = g_j(x)$ for at lest k x's. This implies $g_0 = g_j$, 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

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

That is, it is sufficient to choose q satisfying

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

Let q_0 be the smallest number q satisfying (5). It is not hard to obtain the following estimation of q_0.

Lemma 3

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

Moreover,

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

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

Now, we present an improvement with the following multiplication theorem.

Theorem 4 *If there exist a q-nary $(d, 1; z)$-disjunct $f \times n$ matrix M_1 and a q'-nary $(d, 1; z')$-disjunct $f' \times q$ matrix M_2, then there exists a q'-nary $(d, 1; zz')$-disjunct $ff' \times n$ matrix M_3.*

Proof M_3 can be constructed from M_1 and M_2 by labeling columns of M_2 with $0, 1, \ldots, q-1$ and replacing each entry of M_1 by a corresponding column of M_2.

Consider $d+1$ columns $C_0, C_1, \ldots, C_d$ of M_3. They are obtained from $d+1$ columns $C'_0, C'_1, \ldots, C'_d$ of M_1, respectively. Since M_1 is $(d, 1; z)$-disjunct, there exist z rows of M_1 at each of which the entry a_0 of C'_0 is different from entries $a_1, \ldots, a_d$ of $C'_1, \ldots, C'_d$. Suppose $C''_0, \ldots, C''_d$ are columns of M_2, corresponding $a_0, \ldots, a_d$, respectively. Then, C''_0 is different from $C''_1, \ldots, C''_{d'}$. Since M_2 is $(d, 1; z')$-disjunct, C''_0 has zz' rows at each of which the entry of C''_0 is different from entries of $C''_1, \ldots, C''_{d'}$. Therefore, M_3 has at least zz' rows at each of which the entry of C_0 is different from entries of $C_1, \ldots, C_d$. □

If $f'q' < q$, then $fq < ff'q'$, that is, M_3 gives a transversal design with less number of tests than M_1 does. This multiplication theorem gives a trade-off between the number of tests and the number of families. The former is reduced through increasing the latter.

Let us consider case $z = z' = 1$. By Theorem 2 and Lemma 3, we know that there exist a $(d, 1; 1)$-disjunct q-nary $f \times n$ matrix M_1 such that

$$f = (1 + o(1))\frac{d \log_2 n)}{\log_2(d \log_2 n)},$$

and a $(d, 1; 1)$-disjunct q'-nary $f' \times q$ matrix M_2 such that

$$f' = (1 + o(1))\frac{d \log_2 q)}{\log_2(d \log_2 q)},$$

where q and q' are prime powers satisfying $f \le q \le 2f$ and $f' \le q' \le 2f'$. Note that

$$\frac{d \log_2 n)}{\log_2(d \log_2 n)}$$

is increasing with respect to n and $q < d \log n$ for sufficiently large n. Therefore, for sufficiently large n,

$$ff' = (1 + o(1))\frac{d^2 \log_2 n}{\log_2(d \log_2(d \log_2 n))}.$$

As we apply multiplication theorem more times, the oder of d will increase. Therefore, it is not clear that this method would lead to a construction to approach to the lower bound for the number of tests in a nonadaptive group testing.

References

Du D-Z, Hwang FK, Wu W, Znati T (2006) A new construction of transversal designs. J Comput Biol 13:990–995

D'ychkov AG, Macula AJ, Torney DC, Vilenkin PA (2001) Two models of nonadaptive group testing for designing screening experiments. In: Proceedings of the 6th International workshop on model-oriented designs and analysis, pp 63–75

Farach M, Kannan S, Knill E, Muthukrishnan S (1997) Group testing problem with sequences in experimental molecular biology. In: Proceedings of the compression and complexity of sequences, pp 357–367

Marathe MV, Percus AG, Torney DC (2000) Combinatorial optimization in biology. Manuscript