J Comb Optim (2006) 12:297–301
DOI 10.1007/s10878-006-9634-z

Construction of *d(H)*-disjunct matrix for group testing in hypergraphs

Hong Gao · F. K. Hwang · My T. Thai · Weili Wu · Taieb Znati

Published online: 14 August 2006

Abstract Given a hypergraph with at most d positive edges, identify all positive edges with the minimum number of tests each of which tests on a subset of nodes, called a pool, and the outcome is positive if and only if the pool contains a positive edge. This problem is called the group testing in hypergraphs, which has been found to have many applications in molecular biology, such as the interactions between bait proteins and prey proteins, the complexes of eukaryotic DNA transcription and RNA translation. In this paper, we present a general construction for constructions of nonadaptive algorithms for group testing in hypergraphs.

Keywords Group testing · Pooling designs · Complex · DNA library screening

Weili Wu: Support in part by National Science Foundation under grant ACI-0305567.

Taieb Znati: Support in part by National Science Foundation under grant CCF-0548895.

H. Gao
Department of Computer Science Harbin Institute of Technology Harbin 150001, P. R. China
e-mail: gaohong@hit.edu.cn

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

M. T. Thai
Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, USA
e-mail: mythai@cise.ufl.edu

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

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

1. Introduction

Given a set of n items with at most d positive ones, identify all positive items with less number of tests; each test is on a subset of items, called a *pool*, and the test outcome has two possibilities, *positive* and *negative*. The positive outcome means that the pool contains at least one positive item and the negative outcome means that the pool does not contain any positive item. This problem is called *group testing*. The group testing has been studied since 1943 (Dorfman, 1943). It has applications in many areas, such as medical testing, multi-channel, computer networks (Du and Hwang, 2006), and especially in molecular biology, e.g., DNA library screening (D'ychkov et al., 2001; Farach et al., 1997; Wu et al., 2004), physical mapping, contig sequencing, and gene detection (Du and Hwang, 2006). Motivated from some applications in molecular biology, a new model, group testing in complex have been promoted and has been studied extensively (Macula et al., 2000, 2004; Torney, 1999; Triesch, 1996).

In the complex model of group testing, the positive outcome of a testing on a pool is usually due to the combination effect of several items rather than an individual item. That is, given n items and a collection of at most d positive subsets, the problem is to identify all positive subsets with less number of tests. Each test is on a pool. The test outcome is positive if and only if the pool contains a positive subset.

The group testing in complex is a special case of the group testing in hypergraph. The latter is as follows: Given a hpergraph H with at most d positive edges, identify all positive edges with less number of tests. Each test is on a pool, i.e., a subset of nodes. The test outcome is positive if and only if the sub-hypergraph induced by the pool contains a positive edge. In the complex model of group testing, all items form the node set and all suspected subsets of nodes form the edge set.

An algorithm for group testing is *nonadaptive* if all tests are arranged in a single round, that is, no information on test outcomes is available for determining the pool of another test. It has been very well-known that compared with sequential group testing, the nonadaptive group testing usually takes a short time with a little more number of tests. However, for applications in molecular biology, nonadaptive group testing is promoted due to the time-consuming of each test.

An algorithm for nonadaptive group testing can be represented by a binary incidence matrix. Its columns are labeled with all vertices and its rows are labeled with all pools. The cell (i, j) contains an 1-entry if and only if the ith pool contains the jth vertex. The binary incidence matrix M of a nonadaptive algorithm for group testing in a hypergraph H is $d(H)$-*disjunct* if for any $d+1$ edges $E_0, E_1, \ldots, E_d$ of H, there exists a row, or say a pool, containing E_0, but not $E_1, \ldots, E_d$. A $d(H)$-disjunct matrix can identify all positive edges in a sample with at most d positive edges in a very simple way that an edge is negative if and only if it is contained in a negative pool.

There exist several constructions of nonadaptive algorithms for group testing in hypergraph (Du and Hwang, 2006). Each construction is usually for a certain class of hypergraphs. In this paper we give a general construction of $d(H)$-disjunct matrix for any hypergraph H by extending a construction of Du et al. (2004) for transversal designs.

2. Main results

A binary matrix M is said to be $(d;c)(H)$-*disjunct* for a hypergraph H if M is the incidence matrix of H and for any $d+1$ edges $E_0, E_1, \ldots, E_d$ of H, there exists at least c rows, or say c pools, each containing E_0, but not $E_1, \ldots, E_d$. Clearly, the $(d;1)(H)$-disjunctness is equivalent to the $d(H)$-disjunctness. A $(d;c)(H)$-disjunct matrix can be used to identify d positive edges even if up to $c-1$ test outcomes contain errors. To seek generality, We will construct $(d;c)(H)$-disjunct matrices instead of $d(H)$-disjunct matrices.

Consider a hypergraph $H=(V,\mathcal{E})$ satisfying condition that there do not exist two edges E and E' in H such that $E \subset E'$ or $E' \subset E$. Let $GF(q)$ be a finite field of order q. For each vertex $v \in V$, we associate it with a polynomial p_v of degree $k-1$ over $GF(q)$. Thus, each edge E in $\mathcal{E}$ would associate with a subset of polynomials of degree $k-1$ over $GF(q)$, $P_E = \{p_v \mid v \in E\}$.

Let S be a subset of s elements in $GF(q)$. First, we construct a $s \times |\mathcal{E}|$ matrix $A_H(q,k,s)$ with row labels in S and column labels in $\mathcal{E}$. Each cell (x, E) contains a subset of elements in $GF(q)$, $\{p_v(x) \mid vinE\}$.

Theorem 1. *Let r denote the maximum cardinality of an edge in $\mathcal{E}$. Suppose $s \geq rd(k-1)+c$. Then $A_H(q,k,s)$ has the property that for any $d+1$ columns $C_0, C_1, \ldots, C_d$, there exists at least c rows at each of which the entry of C_0 does not contain the entry of C_j for all $j = 1, 2, \ldots, d$.*

Proof: Suppose to the contrary that such c rows do not exist. Then among any c rows, there exists a row such that the entry of C_0 contains the entry of C_j for some $j \in \{1, 2, \ldots, d\}$. This means that there exist at least $rd(k-1)+1$ rows at each of which the entry of C_0 contains the entry of C_j for some $j \in \{1, 2, \ldots, d\}$. Thus, there exists a $j \in \{1, 2, \ldots, d\}$ such that entries of C_0 contain corresponding entries of C_j at least $r(k-1)+1$ rows. Let E_0 and E_j be edges associated with columns C_0 and C_j, respectively. Then $P_{E_0}(x) \supseteq P_{E_j}(x)$ for at least $r(k-1)+1$ distinct values of x where $P_E(x) = \{p(x) \mid p \in P_E\}$. Since $|E_0| \leq r$, for any $u \in E_j$, there exists $v \in E_0$ such that $p_u(x) = p_v(x)$ for at least k distinct values of x. It follows $p_u = p_v$. Hence, $P_{E_0} \supseteq P_{E_j}$. This means that $E_0 \supseteq E_j$, contradicting our assumption on H. □

Now, we construct a $d(H)$-disjunct matrix $B_H(q,k,s)$ from $A_H(q,k,s)$. $B_H(q,k,s)$ has $|V|$ columns labeled with all nodes of H. For each row x of $A_H(q,k,s)$ and each entry F in row x, we construct a row with label $< x, F >$ for $B_H(q,k,s)$ as follows: Put an 1-entry in cell $(< x, F >, v)$ if $p_v(x) \in F$, and put a 0-entry in cell $(< x, F >, v)$, otherwise.

Theorem 2. *Let r denote the maximum cardnality of an edge in $\mathcal{E}$. Suppose $s \geq rd(k-1)+c$. Then $B_H(q,k,s)$ is $(d;c)(H)$-disjunct.*

Proof: Consider $d+1$ edges $E_0, E_1, \ldots, E_d$ of H. By Theorem 1, $A_H(q,k,s)$ has c rows $x_1, x_2, \ldots, x_c$ such that the entry F in cell (x_i, E_0) does not contain the entry at cell (x_i, E_j) for all $j = 1, 2, \ldots, d$. This means that the row $< x_i, F >$ of $B_H(q,k,t)$

corresponds to a pool which contains E_0, but not E_j for all $j = 1, 2, \ldots, d$. Therefore, $B_H(q, k, s)$ is $(d; c)(H)$-disjunct. □

Suppose H has m edges and n vertices. For existence of $A_H(q, k, s)$, q, k, s and n must satisfy the following conditions:

$$s \leq q \tag{1}$$

since there must exist at least s row labels, and

$$n \leq q^k \tag{2}$$

since there must exist at least n column labels. By Theorem 2, for $B_H(q, k, s)$ to have $(d; c)(H)$-disjunctness, it suffices to have

$$s \geq dr(k-1) + c. \tag{3}$$

From (1), (2) and (3), we can obtain the following

Theorem 3. *There exists a $(d; c)(H)$-disjunct matrix $B_H(q, k, t)$ with*

$$q \leq (2 + o(1)) \frac{d \log_2 n}{\log_2(d \log_2 n)}.$$

Moreover, $B_H(q, k, t)$ has at most $q(q+1)^r$ rows.

Proof: The proof is similar to an analysis in Du et al. (2004). □

References

Barillot E, Lacroix B, Cohen D (1991) Theoretical analysis of library screening using a N-dimensional pooling designs. Nucleic Acids Res 19:6241–6247

Dorfman R (1943) The detection of defective members of large populations. Ann Math Statist 14:436–440

Du D-Z, Hwang FK (1999) Combinatorial group testing and its applications (2nd ed.). World Scientific, Singapore

Du D-Z, Hwang FK (2006) Pooling designs: Group testing in molecular biology. World Scientific, Singapore

Du D-Z, Hwang FK, Wu W, Znati T (2004) A new construction of transversal designs, manuscript

D'ychkov AG, Macula AJ, Torney DC, Vilenkin PA (2001) Two models of nonadaptive group testing for designing screening experiments. In: Proc. 6th Int. Workshop on Model-Orented 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: Proc. Compression and Complexity of Sequences, pp 357–367

Ngo HQ, Du D-Z (2000) A survey on combinatorial group testing algorithms with applications to DNA library screening. In: Discrete mathematical problems with medical applications (New Brunswick, NJ, 1999), pp. 171–182, DIMACS Ser. Discrete Math Theoret Comput Sci, 55, Amer Math Soc, Providence, RI

Macula AJ, Torney DC, Villenkin PA (2000) Two-stage group testing for complexes in the presence of errors. DIMACS Sries Disc Math Theor Comput Sci 55:145–157

Macula AJ, Rykov VV, Yekhanin S (2004) Trivial two-stage group testing for complexes using almost disjunct matrices. Disc Appl Math 137:97–107

Marathe MV, Percus AG, Torney DC (2000) Combinatorial optimization in biology, manuscript
Park H, Wu W, Wu X, Zhao HG (2003) DNA screening, nonadaptive group testing, and simplicial complex, J Comb Optim 7:389–394
Thierry-Mieg N, Trilling L, Roch J-L (2004) Anovel pooling design for protein-protein interaction mapping. manuscript
Torney DC (1999) Sets pooling designs. Ann Comb 3:95–101
Triesch E (1996) A group testing problem for hypergraphs of bounded rank. Disc Appl Math 66:185–188
Wu W, Li C, Huang X, Li Y (2004) On error-tolerant DNA screening. Discrete Applied Mathematics