Optimization Letters (2008) 2:425–431
DOI 10.1007/s11590-007-0070-5

ORIGINAL PAPER

An upper bound of the number of tests in pooling designs for the error-tolerant complex model

Hong-Bin Chen · Hung-Lin Fu · Frank K. Hwang

Received: 9 June 2006 / Accepted: 1 November 2007 / Published online: 28 November 2007

Abstract Recently pooling designs have been used in screening experiments in molecular biology. In some applications, the property to be screened is defined on subsets of items, instead of on individual items. Such a model is usually referred to as the complex model. In this paper we give an upper bound of the number of tests required in a pooling design for the complex model (with given design parameters) where experimental errors are allowed.

Keywords Pooling designs · Nonadaptive algorithms · Disjunct matrices

1 Introduction

In the classical group testing problem (see [3,5] for surveys), we consider a set N of n items consisting of at most d defective items with the others being good items. The problem is to identify all defective items with a small number of group tests. A group test can be applied to an arbitrary subset $S \subseteq N$ with two possible outcomes; a negative outcome implies all items in S are good while a positive outcome implies otherwise, i.e., there exists at least one defective item in S, not knowing which or how many.

Group testing is a basic tool which can be applied to a variety of problems such as blood testing, multiple access communication, coding theory, among others [5]. Group testing procedures have been recently applied to computational molecular biology. Recent advances in molecular biology, especially the success of the Human Genome Project, have made the study of gene functions more popular. The study of gene functions requires a high quality DNA library, which is a collection of the copies of DNA

H.-B. Chen (✉) · H.-L. Fu · F. K. Hwang
Department of Applied Mathematics, National Chiao Tung University, Hsinchu 30050, Taiwan
e-mail: andan.am92g@nctu.edu.tw

segments called *clones*. In screening a clone library, the goal is to determine which clones in the clone library hybridize with a given probe in an efficient fashion. A clone is said to be positive if it hybridizes with the probe, and negative otherwise. Clearly, classic group testing can be applied to the screening situation. Over the past few years, a considerable number of studies have been made on this topic. For a general reference between group testing and molecular biology, the readers may refer to the book [6].

In molecular biology, the biological objects (e.g., clones, cells, molecules) are being identified, but it remains a challenge to understand how they cooperate to produce various attributes. Torney [17] first introduced a generalized group testing problem geared to such a need. The problem is described as follows. Consider a set N of n molecules and a given family $\mathcal{C}$ of subsets of N. The goal is to identify an unknown family $D = \{D_i\}$ from the given family $\mathcal{C}$, where the joint appearance of all molecules in each D_i causes a certain given biological phenomenon. An experiment, sometimes called a *pool*, can be applied to an arbitrary subset $S \subseteq N$ with two possible outcomes; a negative outcome implies S does not contain any $D_i \in D$, and a positive outcome implies otherwise. A set of molecules which is a member of $\mathcal{C}$ is called a complex; members of D are called positive complexes and the others are called negative complexes. Such a model is usually referred to as the complex model. Of particular note is the basic assumption that members of $\mathcal{C}$ are subject to non-inclusion. The reason for this is not difficult to grasp: it is that no positive complex may include any other positive complex.

Treating each molecule as a vertex and every complex as a hyper-edge, then $\mathcal{C}$ can be viewed as a hypergraph on the vertex set N. Thus, the complex model can be easily fitted into the framework of graph testing, learning the hidden subgraph D in the given hypergraph $\mathcal{C}$ where the only allowed operation is to query whether a set of vertices induces a hyper-edge of D. There are different graph testing problems according to prior knowledge of D; the usual assumption is D has at most d edges, but it can also be D is a matching [1,2] or a hamiltonian circuit [9]. The complex model is also related to other problems such as superimposed codes and secure key distribution, among others [1,4,7,12,17]. Besides, some probabilistic analysis can be found in [13,14].

In the applications to molecular biology, an experiment corresponding to a group test could take several hours or even several days. Thus, it is impractical to perform the experiments sequentially and great importance is attached to *nonadaptive group testing algorithms*, also called *pooling designs* in the literature, in which all experiments are performed simultaneously. For this purpose, one must decide exactly which pools to test before any testing occurs. Another feature of biological experiments is that errors almost always occur during the testing procedure. In practice, the decoding issue becomes even more difficult due to experimental errors, and thus algorithms with error-tolerant ability are desirable. In this paper we focus on the construction of error-tolerant pooling designs for the complex model.

2 Preliminaries

A pooling design can be represented by an incidence matrix M where rows are labeled by tests, columns by molecules, and cell (i, j) has a 1-entry if and only if molecule

j is in test i. For convenience, we view each column as the set of row indices where it has 1-entries. Then we can talk about the union and the intersection of columns. We say that a set X of columns appears (or is contained) in a row means all columns in X have a 1-entry in the row. A pool with a negative (positive) outcome is called a negative (positive) pool, respectively.

A main property of M to solve the group testing problem on complexes, i.e., the $(d, r; z]$-disjunctness, has been discussed in [4,7,10,15,16]. A binary matrix M is said to be $(d, r; z]$-disjunct if for any $d + r$ columns $C_1, C_2, \ldots, C_{d+r}$,

$$\left| \bigcap_{i=1}^{r} C_i \setminus \bigcup_{i=r+1}^{d+r} C_i \right| \geq z.$$

That means for any $d + r$ columns there exist at least z rows where each of the r designated columns has 1-entries and each of the other d columns has 0-entries.

Suppose our only knowledge of D is $|D| \leq d$ and every complex consists of at most r molecules. Then a $(d, r; z]$-disjunct matrix M can be used to identify D if the number of error-tests is at most $\lfloor \frac{z-1}{2} \rfloor$. Consider a negative complex Q and a set $D = \{D_i\}$ of positive complexes, $|D| \leq d$. According to the non-inclusion assumption, we can choose two disjoint sets of columns, an r-set R and a d-set S with $S \cap R = \emptyset$, such that $Q \subseteq R$ and $S \cap D_i \neq \emptyset$ for all D_i. By the $(d, r; z]$-disjunctness property, M has at least z rows each containing R but none of S. The pools corresponding to these rows must test negative since they do not contain any positive complex. Even for the worst case that $\lfloor \frac{z-1}{2} \rfloor$ outcomes are erroneous, Q still appears in at least $z - \lfloor \frac{z-1}{2} \rfloor = \lceil \frac{z-1}{2} \rceil + 1$ negative pools. On the other hand, each positive complex appears in at most $\lfloor \frac{z-1}{2} \rfloor$ negative pools (due to errors). Hence, for each complex, by simply counting the number of negative pools it appears, we can determine whether it is positive or not.

Let $t(n, d, r; z]$ denote the minimum number of rows among all $(d, r; z]$-disjunct matrices with n columns. In this paper, we are interested in providing a method to construct the $(d, r; z]$-disjunct matrices and then to obtain an upper bound of $t(n, d, r; z]$. Our result is obtained by translating the problem into a hypergraph problem. Engel [8] first observed the equivalence between a $(d, r; 1]$-disjunct matrix and a cover of a properly defined hypergraph. Stinson and Wei [15] generalized the equivalence to $(d, r; z]$-disjunct matrices for general z, but used the equivalence only to derive a lower bound of $t(n, d, r; z]$. The hypergraph we construct in this paper is similar to that of Stinson and Wei except that we take a weight-l binary vector as a vertex if and only if $l = w$, while Stinson and Wei relaxed the condition $l = w$ to $r \leq l \leq n - d$. This crucial step allows us to use the Lovász lemma on minimum cover to derive an upper bound of $t(n, d, r; z]$.

Stinson and Wei proved a lower bound $t(n, d, r; z] \geq 0.7c\frac{(d+r)\binom{d+r}{r}}{\log \binom{d+r}{r}} \log n + \frac{c(z-1)}{2}\binom{d+r}{r}$ when n is sufficient large, where c is a constant. Also, they provided two asymptotic upper bounds for $t(n, d, r; z]$ by using some other structures, one is $O\left(z\binom{d+r}{r}(dr)^{\log^* n} \log n\right)$, where the function $\log^*$ is defined recursively by

$\log^*(1) = 1$ and $\log^* n = \log^*(\lceil \log n \rceil) + 1$ if $n > 1$, and the other is $O\left(z\binom{d+r}{r} \log n\right)$. However, we believe that there is a flaw in the latter one, and we will explain the matter later. In the next section, we show that $t(n, d, r; z] < z(k/r)^r (k/d)^d [1 + k(1 + \ln(n/k + 1))]$, where $k = d + r$. Finally, we conclude this paper with some remarks.

3 The main results

Given a finite set X, a hypergraph $H = (X, \mathcal{F})$ is a family $\mathcal{F} = \{E_1, E_2, \ldots, E_m\}$ of subsets of X. The elements of X are called vertices, and the sets E_i's are the edges of the hypergraph H. A hypergraph with $|E_i| = |E_j|$ for all $i \neq j$ is said to be *uniform*. For $u \in X$, define the degree $d_H(u)$ of u to be the number of edges containing u. A hypergraph H in which all vertices have the same degree is said to be *regular*, i.e., $d_H(u) = d_H(v)$ for all $u, v \in X$. A z-cover of H is a subset $C \subseteq X$ such that $|C \cap E_i| \geq z$ for all i. Let $\tau_z(H)$ denote the minimum size among all z-covers of H. It is easy to see that

$$\tau_z(H) \leq z\tau_1(H). \tag{1}$$

By a greedy strategy, i.e., choosing vertices sequentially in X such that every chosen vertex intersects the maximum number of edges which are not covered yet, a fundamental result by Lovász [11] implies that

$$\tau_1(H) < \frac{|X|}{\min_{E \in \mathcal{F}} |E|}(1 + \ln \Delta), \tag{2}$$

where $\Delta = \max_{u \in X} d_H(u)$.

Our aim is to show that $(d, r; z]$-disjunct matrices can be obtained from z-covers of properly defined hypergraphs, and then Lovász's result (2) provides us a desired upper bound of $t(n, d, r; z]$.

Let X_w be the set of all binary vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of length n containing w 1's. For any two disjoint subsets D, R of $[n]$, where $[n]$ denotes the set $\{1, 2, \ldots, n\}$, define $E_{D,R} = \{\mathbf{x} = (x_1, x_2, \ldots, x_n) \in X_w : x_i = 1,\ x_j = 0 \text{ for } i \in R,\ j \in D\}$. For example, $E_{\{1\},\{3\}} = \{(1, 1, 1, 0, 0), (1, 0, 1, 1, 0), (1, 0, 1, 0, 1)\}$ when $w = 3, n = 5$. Then, for $r \leq w \leq n - d$, define the hypergraph $H = (X_w, \mathcal{F})$, where $\mathcal{F} = \{E_{D,R} : |D| = d,\ |R| = r$, and $D \cap R = \emptyset\}$. Note that $|X_w| = \binom{n}{w}$ and $|\mathcal{F}| = \binom{n}{r}\binom{n-r}{d}$.

By the construction, a $(d, r; z]$-disjunct matrix with n columns can be obtained from a z-cover of $H = (X_w, \mathcal{F})$ by treating $\{1, 2, \ldots, n\}$ as the set of columns, each vertex in the z-cover as a row, and the jth column has a 1-entry in that row if and only if $x_j = 1$ in that vertex. The reason is as follows. For any two disjoint d-set D and r-set R of $[n]$, there exists a corresponding edge $E_{D,R}$ in the hypergraph $H = (X_w, \mathcal{F})$. Thus, in the z-cover there are at least z vertices, each of which intersects the edge $E_{D,R}$. According to our transformation, each row corresponding to one of these vertices in the z-cover has a 1-entry in every column of R and a 0-entry in every column of D. In other words, for any $d + r$ columns there exist at least z rows where each of

the r designated columns has 1-entries and each of the other d columns has 0-entries; hence it is a $(d, r; z]$-disjunct matrix. We then obtain the following theorem.

Theorem 3.1 *For any positive integers d, r, w, z and n, with $r \le w \le n - d$, there exists a $t \times n$ $(d, r; z]$-disjunct matrix with*

$$t < \frac{z\binom{n}{r}\binom{n-r}{d}}{\binom{w}{r}\binom{n-w}{d}}\left[1 + \ln\binom{w}{r}\binom{n-w}{d}\right].$$

Proof By the construction of $H = (X_w, \mathcal{F})$, H is uniform and regular; hence $\frac{|X|}{\min_{E\in\mathcal{F}}|E|} = \frac{|\mathcal{F}|}{\Delta}$. Moreover, we have $|\mathcal{F}| = \binom{n}{r}\binom{n-r}{d}$ and $\Delta = \binom{w}{r}\binom{n-w}{d}$. The theorem follows directly from (1) and (2). □

From Theorem 3.1, all we need to do is to minimize the function by properly choosing w to obtain a better bound. Let $f(w) = \binom{w}{r}\binom{n-w}{d}$. Then $f(w+1) = \left(\frac{w+1}{w-r+1}\cdot\frac{n-w-d}{n-w}\right) f(w)$. When $r \le w < n$, clearly the function $\left(\frac{w+1}{w-r+1}\cdot\frac{n-w-d}{n-w}\right)$ is decreasing since the two terms are both decreasing. Accordingly, if we ignore the fact that w must be an integer, then setting $w = \frac{nr-d}{d+r}$, satisfying $\left(\frac{w+1}{w-r+1}\cdot\frac{n-w-d}{n-w}\right) = 1$, maximizes the function $f(w) = \binom{w}{r}\binom{n-w}{d}$; hence in some sense be a good choice to minimize the function in Theorem 3.1. For convenience, denote $k = d + r$.

Theorem 3.2 *For any positive integers d, r, z and n with $d + r \le n$,*

$$t(n, d, r; z] < z(k/r)^r(k/d)^d[1 + k(1 + \ln(n/k + 1))].$$

Proof For given positive integers d, r, z and n, setting $w = n'r/k$ where $n' \ge n$ is the least integer such that $n'r/k$ is an integer, we have

$$\begin{aligned}\frac{z\binom{n'}{r}\binom{n'-r}{d}}{\binom{w}{r}\binom{n-w}{d}} &\le \frac{zn'\cdot(n'-1)\cdots(n'-k+1)}{[(n'r/k)\cdots(n'r/k-r+1)][(n'd/k)\cdots(n'd/k-d+1)]}\\ &\le z(k/r)^r(k/d)^d.\end{aligned}$$

Moreover, using the inequality $\binom{a}{b} \le (ea/b)^b$, where $e \approx 2.7182$ is the base of the natural logarithm, one concludes

$$\binom{n'r/k}{r}\binom{n'-n'r/k}{d} \le e^k(n'/k)^k.$$

From the above inequalities and Theorem 3.1, we have

$$t(n', d, r; z] < \frac{z\binom{n'}{r}\binom{n'-r}{d}}{\binom{w}{r}\binom{n'-w}{d}}\left[1 + \ln\binom{w}{r}\binom{n'-w}{d}\right]$$
$$< z(k/r)^r(k/d)^d[1 + k(1 + \ln(n'/k))].$$

Note that $n' < n+k$ because of the choice of n'. For any given positive integers d, r, z and n, we have

$$t(n, d, r; z] \le t(n', d, r; z]$$
$$< z(k/r)^r(k/d)^d[1 + k(1 + \ln(n'/k))]$$
$$= z(k/r)^r(k/d)^d[1 + k(1 + \ln(n/k + 1))].$$

□

4 Conclusions

What we concerned in this paper is the construction of pooling designs for the complex model. Although a number of studies have been made, little is known about the construction of pooling designs for the complex model. A main design to solve the group testing problem on complexes is $(d, r; z]$-disjunct matrices. In this paper we provide an explicit construction of $(d, r; z]$-disjunct matrices by using Lovász's result (2). Precisely, we show that for any positive integers d, r, z and n with $d+r \le n$, there exists a $t \times n$ $(d, r; z]$-disjunct matrix with $t < z(k/r)^r(k/d)^d[1+k(1+\ln(n/k+1))]$, where $k = d + r$. Our result presented in this paper is the first nontrivial and non-asymptotic bound of $t(n, d, r; z]$. Note that the two upper bounds proposed in [15] by Stinson and Wei are asymptotic.

It is worth pointing out that we believe that the original analysis in [15] has a flaw. The authors showed that for any positive integers d, r, z and n, there exists a $t \times n(d, r; z]$-disjunct matrix with $t = O\left(z\binom{d+r}{r}\log n\right)$, by using a construction of perfect hash families described in [18]. However, the asymptotic result in [18] cited by Stinson and Wei should not be $O(\log n)$, but $O(C \log n)$, where C depends on the parameters d and r (hence it cannot be ignored even under O-notation). In addition, the same flaw can also be found in the construction of $(d, 1; 1]$-disjunct matrices in [18].

Acknowledgments The authors would like to thank to the anonymous referees for careful reading and valuable suggestions.

References

1. Alon, N., Beigel, R., Kasif, S., Rudich, S., Sudakov, B.: Learning a hidden matching. SIAM J. Comput. **33**, 487–501 (2004)
2. Beigel, R., Alon, N., Apaydin, M.S., Fortnow, L., Kasif, S.: An optimal procedure for gap closing in whole genome shotgun sequencing. In: Proceedings of 2001 RECOMB, pp. 22–30. ACM, New York (2001)

3. Balding, D.J., Bruno, W.J., Knill, E., Torney, D.C.: A comparative survey of nonadaptive pooling designs. In: Genetic Mapping and DNA Sequencing, IMA Volumes in Mathematics and Its Applications, pp. 133–154. Springer, Berlin (1996)
4. Chen, H.B., Du, D.Z., Hwang, F.K.: An unexpected meeting of four seemingly unrelated problems: graph testing, DNA complex screening, superimposed codes and secure key distribution. J. Combin. Optim. **14**, 121–129 (2007)
5. Du, D.Z., Hwang, F.K.: Combinatorial Group Testing and Its Applications, 2nd edn. World Scientific, Singapore (2000)
6. Du, D.Z., Hwang, F.K.: Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing. World Scientific, Singapore (2006)
7. D'yachkov, A.G., Vilenkin, P.A., Macula, A.J., Torney, D.C.: Families of finite sets in which no intersection of ℓ sets is covered by the union of s others. J. Combin. Theory Ser. A **99**, 195–218 (2002)
8. Engel, K.: Interval packing and covering in the boolean lattice. Combin. Prob. Comput. **5**, 373–384 (1996)
9. Grebinski, V., Kucherov, G.: Reconstructing a Hamiltonian cycle by querying the graph: application to DNA physical mapping. Discrete Appl. Math. **88**, 147–165 (1998)
10. Kim, H.K., Lebedev, V.: On optimal superimposed codes. J. Combin. Designs **12**, 79–91 (2004)
11. Lovász, L.: On the ratio of optimal integral and fractional covers. Discrete Math. **13**, 383–390 (1975)
12. Macula, A.J., Popyack, L.J.: A group testing method for finding patterns in data. Discrete Appl. Math. **144**, 149–157 (2004)
13. Macula, A.J., Rykov, V.V., Yekhanin, S.: Trivial two-stage group testing for complexes using almost disjunct matrices. Discrete Appl. Math. **137**, 97–107 (2004)
14. Macula, A.J., Torney, D.C., Vilenkin, P.A.: Two-stage group testing for complexes in the presence of errors. DIMACS Ser. Discrete Math. Theoret. Comput. Sci. **55**, 145–157 (1999)
15. Stinson, D.R., Wei, R.: Generalized cover-free families. Discrete Math. **279**, 463–477 (2004)
16. Stinson, D.R., Wei, R., Zhu, L.: Some new bounds for cover-free families. J. Combin. Theory Ser. A **90**, 224–234 (2000)
17. Torney, D.C.: Sets pooling designs. Ann. Combin. **3**, 95–101 (1999)
18. Wang, H., Xing, C.: Explicit constructions of perfect hash families from algebraic curves over finite fields. J. Combin. Theory Ser. A **93**, 112–124 (2001)